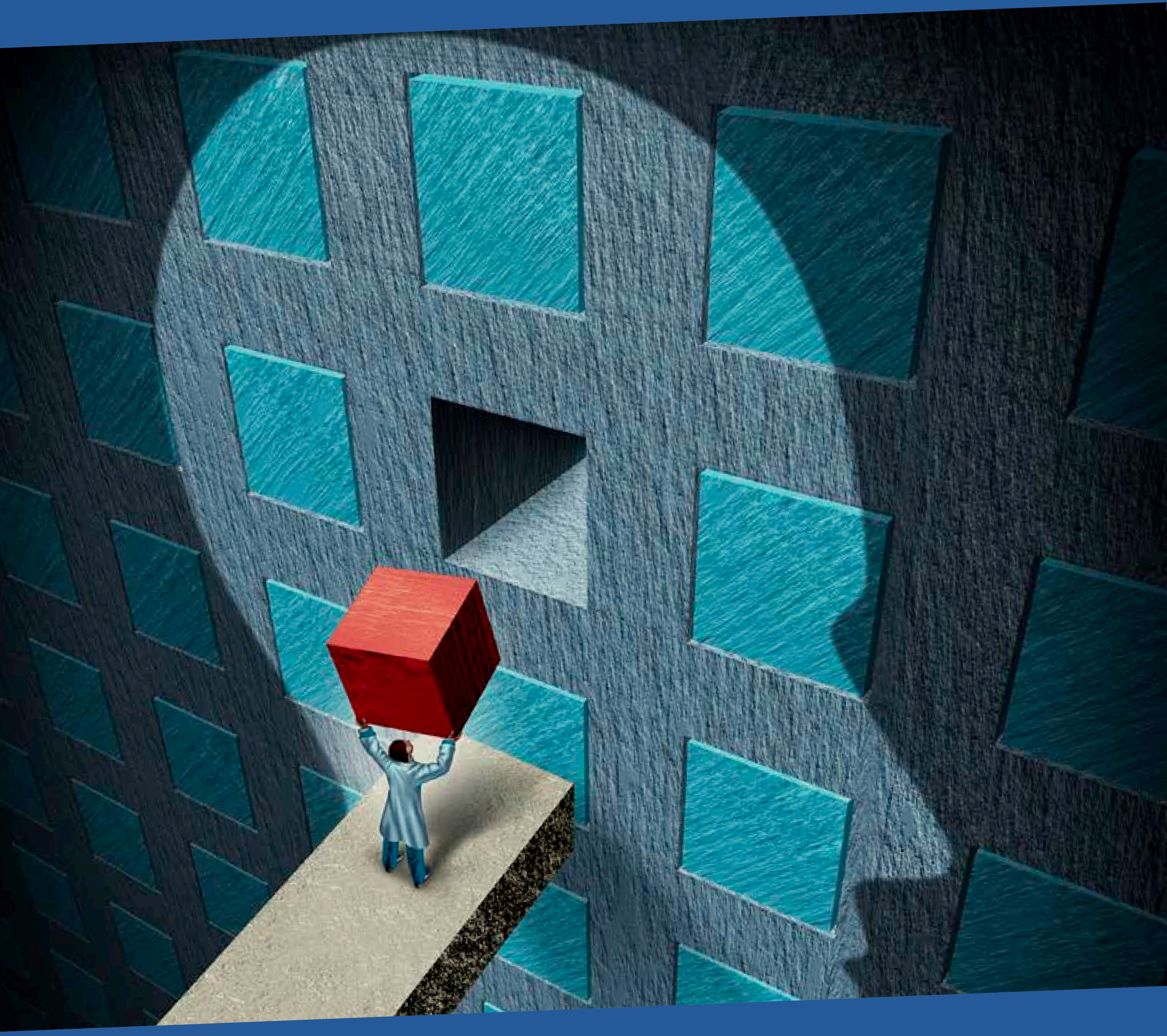
ISBN: 978-9934-564-39-0



RESPONDING TO COGNITIVE SECURITY CHALLENGES

ISBN: 978-9934-564-39-0
Authors: Sebastian Bay, Giorgio Bertolin, Nora Biteniece, Edward H. Christie, Anton Dek, Rolf E. Fredheim, John D. Gallacher, Kateryna Kononova, Tetiana Marchenko
Project manager: Giorgio Bertolin
Text editor: Anna Reynolds, Mike Collier
Design: Kārlis Ulmanis

Riga, January 2019
NATO STRATCOM COE
11b Kalciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook/stratcomcoe
Twitter: @stratcomcoe

This publication does not represent the opinions or policies of NATO.

The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of NATO StratCom COE. NATO StratCom COE does not take responsibility for the views of authors expressed in their articles.

CONTENTS

Introduction.. 4

The Current Digital Arena and its Risks
to Serving Military Personnel...6

A Picture is Worth a Thousand Words:
Analysing Images in the Online Information Environment..................18

The Effects of Banning the Social Network
VK in Ukraine...38

Division Abroad, Cohesion at Home:
How the Russian Troll Factory Works to Divide
Societies Overseas but Spread Pro-Regime Messages at Home.........60

Countering Subversion Online: What Role for Public Policy?...............80

Conclusion...94

01

INTRODUCTION

Giorgio Bertolin

This research product is a collection of different efforts, united by a common goal: to identify some of the most critical security challenges in online environment and what can be done to counter them, and to determine the role of governments and state institutions in countering them. People spend increasing amounts of time online, either communicating, networking, entertaining themselves, or obtaining news. On the one hand, this narrows the number of places analysts must look at when assessing the information environment, on the other, we have yet to fully tap into the analysis potential for this enormous space and leverage it to increase the effectiveness of our communications.

Last year we focused on a single topic, i.e. disinformation. The result was a monograph titled *Digital Hydra: Security Implications of False Information Online*. While disinformation is still a major threat that haunts the information environment, it is by no means the only one. As a broader overview can be extremely beneficial, this volume will consider a number of different issues.

Our exploration begins with an examination of the risks posed by the vast amount of personal information we routinely disseminate online. A number of highly publicised cases during the last year have shown how an actor can collect and use this information in a malicious way to the detriment of users, the armed forces, and societies in general. We explore these risks by studying the foundations of the problem, and by conducting an experiment where we demonstrate how data can be collected and exploited within the context of a military exercise.

The second chapter considers the growing relevance of visual material. Social media analysis is still predominantly text-based, despite the importance and frequency of visual material online. This chapter looks at the tools that are currently available to researchers, and elaborates on a specific case study: visual narratives regarding NATO.

Another chapter is dedicated to the effects of the official ban on the social networking website Vkontakte (VK) in Ukraine. A team of local researchers analysed the issue and discovered that the platform lost audience and popularity as a result of the ban; however, those who stayed active became more connected and started to consume more information from ideological groups. Because of the decrease in users, those who continued to use VK after the ban lost the opportunity to confront themselves with different points of view, therefore political discourse is currently more homogeneous than it used to be.

The following chapter uses quantitative analysis and machine learning to show how an alleged Russian troll factory, the Internet Research Agency (IRA), coordinates messaging across multiple social media platforms and uses real-world events to foster societal tensions abroad. The same agency also uses the same platforms in Russianfor a very different purpose — to spread pro-regime messages among domestic audiences.

The final chapter, written by an analyst from NATO headquarters' Emerging Security Challenges Division, offers an overview of the risks and threats that social media use may pose to liberal democratic systems. This is followed by a discussion on possible future options for public policy that serves as a conclusion for the research product as a whole.

Social media give users the power to spread and receive contaminated information.

Threats to cognitive security should not be overlooked. Technological innovations are used to exacerbate deep-seated weaknesses that can destabilise our societies. We hope this anthology will inform the work of researchers and practitioners alike, refining the capabilities of those who are tasked with the safety of our nations and our Alliance.

02

THE CURRENT DIGITAL ARENA AND ITS RISKS TO SERVING MILITARY PERSONNEL

Sebastian Bay, Nora Biteniece

ABSTRACT

The last few years have provided an abundance of examples of how malicious actors can exploit user data to the detriment of social media users, armed forces, and society. This study explores what kind of user data is available in the digital environment and demonstrates how a malicious actor can exploit this data in the context of a military exercise. The results of an experiment conducted by a NATO StratCom COE research team suggest that in the current digital arena an adversary would be able to collect enough personal data on soldiers to create targeted messages with precision, successfully influencing their chosen target audience to carry out desired behaviours.

INTRODUCTION

In the wake of the Cambridge Analytica scandal, the broad media coverage of Facebook CEO Mark Zuckerberg's appearance before the US Congress, and the implementation of the General Data Protection Regulation (GDPR) in the European Union, the news-watching public is becoming increasingly aware that data is constantly being collected about virtually every aspect of our digital lives.

Whenever we browse the internet, purchase goods online, move around the world with our smartphones, or interact with our peers, we generate large amounts of data that are collected by social media companies, internet service providers, and data brokers. With the advent of Internet of Things (IoT), data is now also being collected about our health, our homes, our pets, as well as about our digital equipment and how we use it. The International Data Corporation (IDC) forecasts that by 2025 annual global data creation will have grown tenfold to 163 zettabytes.[1,2]

Cambridge Analytica allegedly analysed thousands of data points on hundreds of millions of Americans to generate effective microtargeting and behaviour-prediction algorithms during the 2016 US presidential election campaign. In light of these events, it is imperative that we increase our understanding of the possibilities for malicious use of data.[3] Much of the data used by Cambridge Analytica was collected

> ”Data is now also being collected about our health, our homes, our pets, as well as about our digital equipment and how we use it. Malicious use of

using the Facebook app 'This is Your Digital Life'. Roughly 270,000 people used this app and unwittingly shared their personal data, and that of their friends, with Cambridge Analytica. It has been estimated that the personal information of roughly 50 million Americans was harvested this way.[4] And Cambridge Analytica is not the only company collecting data on private citizens. Data has become an important component of our digital existence because people now expect customised search results and an online experience tailored to their personal needs, wants, and desires. This kind of customisation is not possible without extensive data collection.

In his testimony to Congress in April 2018, Facebook CEO Mark Zuckerberg described the data Facebook collects on its users as 'information people choose to share online' and 'data needed to make ads relevant'.[5] However, this description leaves out two kinds of data — the metadata users share involuntarily (online behaviour, personal activities, type of hardware used) and data derived, inferred, or predicted from the data shared and generated by users. These kinds of data can reveal surprising insights about both individuals and groups. In simple terms, information about the things users post and like online, combined with where they are, how they travel, and which devices and apps they use, can be used to make predictions about individuals gender, sexual orientation, political leanings, personality, and other characteristics that define us as people.[6,7] Facebook also uses metadata, such as device model, whether they are using WiFi or have been travelling abroad, to add users to categories advertising clients can use to target the users. Facebook also infers characteristics, such as 'potentially interested in switching mobile carrier', from this metadata. A malicious actor could potentially combine an inferred trait such as 'potentially interested in switching mobile carrier' with provided data such as listed employer e.g., 'Country X Armed Forces' to target specific users much more effectively.

Although many know that their online presence leaves many digital traces, far fewer are aware that by using various combinations of data (such as calls, SMS, Bluetooth, and app usage) researchers have been able to predict users' 'Big Five' personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) to model personality/psychopathology.[8] Indeed, knowledge of any four apps installed on a person's smartphone has proven enough to identify 95% of users in a given data set.[9,10,11]

This report discusses what kind of user data is available in the digital environment, and how a malicious actor might exploit this data. In this report "malicious use of data" refers to the usage of data exploiting vulnerabilities in order to deceive, disrupt, interfere and ultimately do harm to individuals and/or society. In assessing if the usage of data is malicious, we have based our discussions on the DIDI-model[12] proposed by Pammet at al., for diagnosing illegitimate influence.[13]

In this report we also present the results of an experiment conducted by the NATO Strategic Communications Centre of Excellence to discover how a malicious actor might exploit user data within the context of a national military exercise.

What kind of user data is available?

Data brokers, such as Acxiom and Epsilon, have grown into multibillion-dollar companies that make a living out of collecting and reselling data.

Data brokers do one of three things:

1) search for information about individuals (name address, income, debt, family)[14] develop dossiers on individuals (age, demographics, family, interests, contact information, health information, etc.);[15]

2) group individuals into segments marketers can use for targeted advertisements;

3) gather information to verify identities and assess risk (often financial risk).[16] Data brokers combine data they have collected with social media data, i.e., user actions (such as providing gender, age, and location), and user interactions on social media platforms (such as liking posts, joining groups, and using a specific device), to create custom segments or detailed digital portraits of targeted individuals and groups.[17]

Social media companies allow advertisers to target their users by uploading custom audience sets, enabling advertisers to use outside datasets (e.g. subscribers to an email list or a contact list of individuals who have recently bought a certain item) to target ads to individuals on the social media. The resulting list (e.g. people who have bought a certain item and use a specific social media platform) can then be further refined using additional data derived from the social media platforms. The result is a highly customised datasets enabling unprecedented microtargeting of social media users.[18] Recent developments have pushed social media companies to limit the use and abuse of ads targeted based on psychographics visible to individual users (also known as 'dark ads')[19], and third-party data,[20] but so far most companies still allow the use of third-party data for targeting ads on their platforms. Social media companies also use third-party data to track and measure ads and ad engagement criteria, such as sales and sentiment.

We have reached a point where it is no longer possible to have a complete overview of the data we use and generate. The number of data points available on any one individual cannot be counted, as they are created and re-created non-stop. Public government data, social media data, and commercial data, together with data aggregated and inferred from these records, create enormous amounts of data with unimaginable scope. This does not mean that comprehensive data is available about every individual, but it does mean that ad targeting is gradually becoming more and more precise, creating unprecedented possibilities for the use and abuse of data.

How can data be used in a malicious way?

The malicious use of data is a more serious problem than targeted messaging. The collection and use of personal data for criminal objectives can have consequences that go far beyond influencing the behaviour of potential customers. Below we identify some of these risks associated with data collection and analysis, taking into consideration the information security principles of confidentiality, availability, and integrity.

MALICIOUS WAYS

Manipulation

A diverse range of industries – from finance and insurance to health and migration – collect data to make business decisions. Since most people usually aren't aware this data about them exists, or what kinds of decisions are being made based on this data, there is a risk that inaccurate data could have severe consequences for individuals without their knowledge.[31] For example, inaccurate data could prevent a person from securing a loan or being granted a security clearance. Inaccurate data can cause an organisation to make erroneous decisions and lost data can be difficult or expensive to replace.

Impersonation

Access to personal information makes it easier for malicious actors to impersonate people online. Personal information can also be used to predict passwords and answer security questions in order to gain access to accounts, and to convince companies and government entities to take specific actions. The confidentiality of personal data is essential for the proper functioning of authorisation layers that control access to sensitive information.

Doxing

Doxing is the technique of intentionally releasing selected information about an individual to influence public perception of that individual, or the creation of conditions and vulnerabilities that can be exploited. The confidentiality of information safeguards the credibility of both individuals and organisations.

Sensitive information

Data generated by our devices, particularly by our mobile devices, often reveal sensitive information about the locations and activities of the people using them. During the Russian annexation and occupation of Crimea and Eastern Ukraine, Russian soldiers and civilians shared a wealth of information that made it possible to verify Russian aggression in Ukraine. The open source verification organisation Bellingcat was able to determine precisely how a Buk missile launcher reached a particular field in eastern Ukraine, who organised the transport, where the missile launcher came from before it arrived in Ukraine, and even identify the (near-complete) history of a single launch unit. However, advances in this area also provide opportunities for malicious actors to exploit data leakage, creating new risk in the military domain.

EXPERIMENTATION

A research team from NATO Strategic Communications Centre of Excellence conducted an experiment in support of a military exercise in an Allied country. We embedded a research team within a red-team cell[21] to evaluate how much data we could collect about exercise participants, to test different open-source intelligence techniques, and to determine if we would be able to induce certain behaviours such as leaving their positions, not fulfilling duties, etc. using a range of influence activities based on the acquired data.

The research team collected open source data during a military exercise targeting armed forces personnel. To protect the privacy of those taking part in the military exercise, no personal data identified during the experiment was stored.[22] The experiment focused on the active phase of the military exercise. The preparations made by the research team took three to four weeks; these included planning the operation, setting up the necessary online accounts, assessing the online information environment, and creating a range of messages and lines of persuasion. The scope of the experiment was limited in comparison to large-scale efforts such as the work undertaken by the Kremlin's Internet Research Agency to influence the US presidential election 2016. An operation of that scale requires months of preparation to set up the necessary infrastructure and develop quality target audience analysis.[23]

Methodology

To assess the extent to which we would be able to exploit social media and open source data to gather information on and influence military personnel during a military exercise, the research team used:

- Impersonation[24]
- Honeypot pages[25]
- Social engineering[26]
- General monitoring and befriending of accounts
- Peoples search engines[27] and open source databases

The level of personal information that could be found using the above methods was very detailed and enabled the research teams to craft influence activities. Information about the exercise itself was found both from exercise participants and public sources such as news and official armed forces pages.

We monitored exercise participants using their Facebook, Instagram, and Twitter accounts. These platforms provided the research team with access to basic information about their targets as they allow users to search by name/username and view any information that has been made public by the platform users.

Results and Findings

The methods employed by the research team resulted in honeypot pages and groups being liked and joined by exercise participants. Shortly after the groups were created and promoted, Facebook shut down the honeypot pages,[28] which meant that the audience acquired through Facebook Ads was lost, and researchers could no longer advertise exclusively to followers of the honeypot pages.

The members of the closed groups were used as a starting point to gather more information. As described in the previous section, researchers searched for information about their targets in public sources, monitored their social media accounts, and attempted to engage them directly via group discussions and messages. The exact methods and their success cannot be disclosed due to operation security.

Overall, we identified a significant amount of people taking part in the exercise and managed to identify all members of certain units, pinpoint the exact locations of several battalions, gain knowledge of troop movements to and from exercises, and discover the dates of the active phases of the exercise. The level of personal information we found was very detailed and enabled us to instil undesirable behaviour during the exercise.

We found that Instagram was popular among soldiers during the exercise, and therefore provided the timeliest information. Facebook, by comparison, was a good starting point for identifying individuals and for mapping their links to other members of the armed forces using the suggested friends feature.[29] Twitter was rarely used during the exercise, and gave no useful information.

The soldiers who were targeted using social engineering shared more information with researchers than the information that could be found about them on their social media accounts. We managed to get an approximate location (+/-1km) for exercise participants, including soldiers from high value units, i.e., units that were required to complete a mission. We obtained phone numbers, email addresses, and pictures of equipment from all participants targeted using social engineering.

Social Media Countermeasures

An important part of our experiment involved the creation of honeypot pages, groups, and profiles on social media to gather data and to test the countermeasures of social media companies. During the exercise, we created honeypot Facebook pages that published information from other sources regarding the exercise, and Facebook pages impersonating the official armed forces page. In addition, we created several social media accounts. Four accounts impersonated real people from the armed forces and one account was entirely fake. The social media companies deployed counter measures to counter our abuse of their platforms with varying degrees of success.

The table below summarises the social media countermeasures we experienced during this process:

Type	Uptime	Cause
Honeypot pages	2 weeks	Reported to Facebook
Pages impersonating existing page	Suspended after 1–2 hours	Did not comply with Facebook T&C
Closed groups	Never suspended	
Fake profile	Never suspended	
Profiles impersonating real people	From 2 hours to infinite Two profiles suspended after 2 hours One profile suspended after one day One profile was never suspended	Reported to Facebook Suspicious activity detected by Facebook

Social Media Vulnerabilities

Prior to the experiment, we found that Facebook only partially respects the privacy settings for workplace disclosure. Accounts that did not publicly display their workplace, still appeared in results when searching for employees using a certain Facebook feature. The security team at Facebook has been informed about this "bug".

We also noticed several profiles that were clearly fake, or not related to the target country in any way, which listed the armed forces targeted by the research team as their workplace. This is a potential vulnerability that malicious actors can exploit — private accounts are allowed to list any entity as employer, which creates a situation whereby accounts can choose to intercept public information intended only for a certain group. There is no simple solution to this problem, as a new set of security challenges would stem from attempts to ensure that only actual employees are able to declare a particular place of work on their Facebook profiles.

Both of these vulnerabilities underscore one important thing — the privacy features and settings of social media platforms cannot be trusted not to leak information to other layers of the social media platform, or to other users and companies with an interest in such information.

CONCLUSIONS

In an essay entitled Preparing for Elections, Facebook CEO Mark Zuckerberg stated that his focus for 2018 is to defend elections against interference, protect the community from abuse, and make sure individuals have more control of their information.[30] These are all important and complex steps that must to be taken by all responsible and serious actors. After years of social media manipulation by malicious actors, we finally have movement in the right direction.

However, states and its citizens need more than verbal assurances that our vital assets will be protected. We must probe, test, and continuously evaluate how data exploitation by malicious actors can threaten allied goals and interests. We need to build not only an infrastructure that protects us, but also improve the training and exercises that test our ability to detect and counter influence activities.

Our experiment showed that, at the current level of information security, an adversary is able to collect a significant amount of personal data on soldiers participating in a military exercise, and that this data can be used to target messages with precision, successfully influencing members of the target audience to carry out desired behaviours.

However, although we managed to collect data and induce behaviour detrimental to the conduct of military operations, we also faced a number of difficulties indicating that social media companies are increasing their efforts to prevent abuse of their platforms. Facebook in particular provided significant pushback, and several of our fake accounts and pages were suspended during the course of the experiment. The fact that social media abuse has been much debated as a phenomenon during the last year has increased public and institutional awareness of the risks and challenges. The effect of this heightened sensitivity was that several of our fake profiles and pages were reported by the armed forces we targeted, and on one occasion a warning for the fake page we had created was circulated.

Even so, despite heightened sensitivity and active users reporting suspicious behaviour, we were successful on a number of occasions, proving that misuse of social media platforms for targeting purposes is still quite possible. Our experiment showed that much remains to be done to improve security, both by the social media companies and by the armed forces. Some of the flaws that enabled us to manipulate social media and social media users are human flaws that can only be addressed through better training and stricter control. But other flaws, such as the lack of transparency, opportunities for microtargeting, and misuse of anonymity, are vulnerabilities built into the social media platforms themselves; this highlights the continuing need to improve these platforms. Two immediate changes that the social media platforms should consider in order to reduce vulnerabilities are:

- Stricter control of the 'suggested friends' feature — a friend should not be suggested unless the user has accepted the friend request. As it stands now, this feature made it extremely easy for us to map out entire units and battalions by identifying only a single member of a unit.

- Preventing search features to showing hidden data — searches should not be allowed to show results that have intentionally been hidden from the public profile by the users.

Our final conclusion is an old conclusion that bears repeating. The armed forces must step up monitoring and countermeasures to reduce the risk of social media being used to gather mission-sensitive information. This is, and will continue to be, a significant challenge in the years to come.

03

A PICTURE IS WORTH A THOUSAND WORDS: analysing images in the online information environment

Nora Biteniece

ABSTRACT

As social media, and the web as a whole, become more visual, organisations and governments can no longer rely solely on textual analysis when seeking to better understand their audiences in the online environment. Methods for thorough analysis of the online information environment, including visual content, must be developed. This chapter discusses the type of information images contain and how it can be extracted, and proposes a computer-based image retrieval to extract valuable information from large volumes of images and aggregate it in a meaningful way.

1. BACKGROUND

An extensive Daesh campaign calling on Muslims to live in the Caliphate is a good example of how the online information environment can being used to influence young Muslim men and women. Daesh have frequently reported on life in the 'Caliphate' in the group's online magazine *Dabiq*. These stories are illustrated with images depicting a successfully governed state, complete with police, schools, and hospitals. Jihadists from around the world travel to Daesh-occupied territories to live in the 'Caliphate' and fight alongside its adherents.

Activities on social media undertaken in preparation for the battle of Mosul provide another, more aggressive example of Daesh's use of the information environment to intimidate their adversaries – they published photographs and videos showing the brutal execution of Iraqi and Syrian soldiers they had captured on social media. As a result of this online terror campaign, many thousands of people fled Mosul just before the main military operation was launched.[32] Both examples demonstrate Daesh's extensive use of visual content, i.e. videos and images, to deliver their messages.

This increased use and consumption of visual content is a recent shift in the online information environment. Every day over three billion photos are shared on social media.[33] Of the top fifteen social media platforms, ten are purely video and picture sharing services, or services providing in-platform functionality for generating and sharing images and videos. Users increasingly choose to send a 'snap',[34] or to 'Instagram it',[35] instead of writing status updates to share their experiences. As social media, and the web as a whole, become more visual, organisations and governments can no longer rely solely on textual analysis when seeking to better understand their audiences in the online environment. A thorough analysis of the entire online information environment, including visual content, is vital for governments and institutions to gain a better understanding public sentiment, to identify audience segments vulnerable to particular types of messaging, and to identifying disinformation efforts, hostile narratives, and early warning signs of potential hybrid threats.

When analysing visual content, the biggest challenges lie in the sheer volume available online, and in the fact that images are more difficult to analyse quantitatively than qualitatively.[36] Image analysis includes the examination of every picture shared by every member of a defined group, aggregating the meanings of the images for the entire group, and then categorising them so as to provide an overview showing trending topics and potential warning signs. Depending on the size of the target audience, this can be an impossible task. Many organisations take their cues from the commercial sector and use social media listening tools to monitor and analyse the online information environment, including visual content.[37]

The general principle is to gather online content that mentions certain keywords or features logos of interest, and use it to calculate a number of pre-defined metrics. These metrics include 'volumes over time', 'by platform', 'by user', 'topics', 'social groups', 'influencers', and others. However, this approach is optimised for recognising brand logos, and will not recognise relevant visual content that does not feature the narrowly-defined visual references of interest here, so the challenge of gathering and analysing relevant visual material online remains.

This report proposes using a computer-based image retrieval to extract valuable information from large volumes of images and aggregate it in a meaningful way. The next chapter discusses the type of information images contain and how it can be extracted. The third chapter presents a case study for image analysis in the context of online information environment analysis. We have chosen to examine online visual narratives regarding NATO and its presence in the three Baltic States and Poland — four battlegroups were deployed along NATO's Eastern flank following the 2016 Warsaw summit, one in each state. The fourth chapter discusses the pros and cons of the proposed approach and the conclusions drawn from the case study.

2. IMAGE ANALYSIS

Image analysis is a loosely defined term and its meaning varies among diverse fields such as computer graphics, digital signal processing, medical imaging etc. This report examines image retrieval as a method for analysing information about the 'imaged objects' using computer-based techniques. In this context, the process of image analysis consists of several smaller problems that must be solved before an image is 'understood'. For example, recognising objects in an image is a separate problem from recognising characters or emotion, and both need to be solved before everything in the image is fully recognised. However, even this does not guarantee that an image is fully understood — hidden meanings, sarcasm, and other contextual information computers are unable to recognise, may remain undetected. This task becomes even more complicated when the analysis must contribute to our overall understanding of the information environment.

An important distinction to be made is between object recognition and object detection. **Object detection** is primarily concerned with where a specific object is located in a given image. **Object recognition** is primarily concerned with which object/s is/are depicted in the image. At its core, object recognition is a classification problem. To solve it, objects in a given image must be assigned a class. Classification problems are generally solved using machine-learning algorithms.[38] Figure 1 illustrates the basic principle of machine learning.

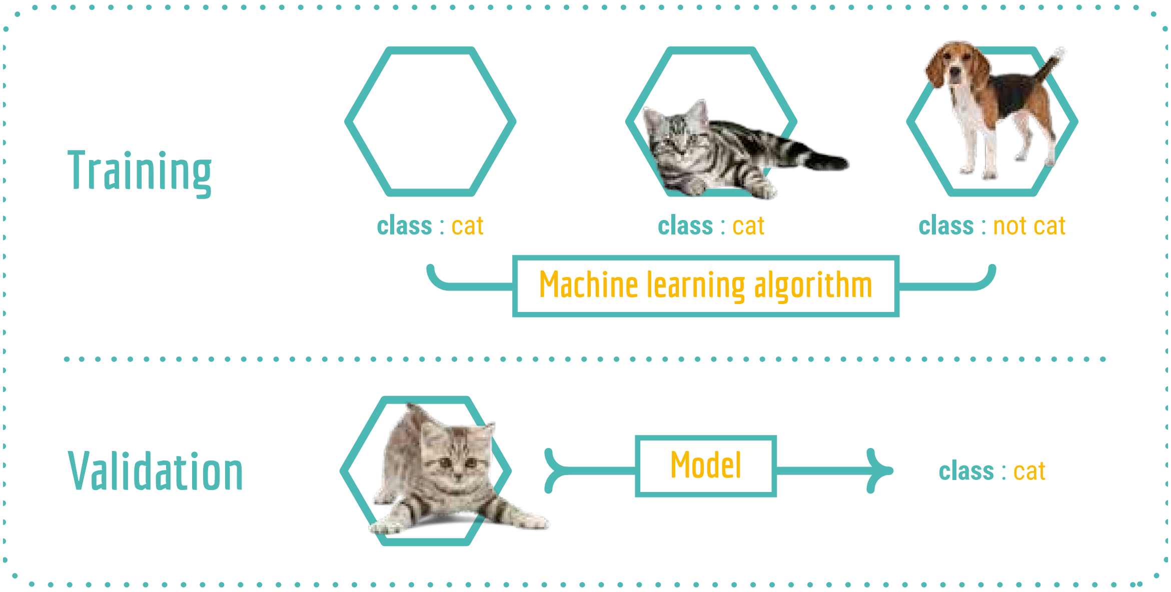
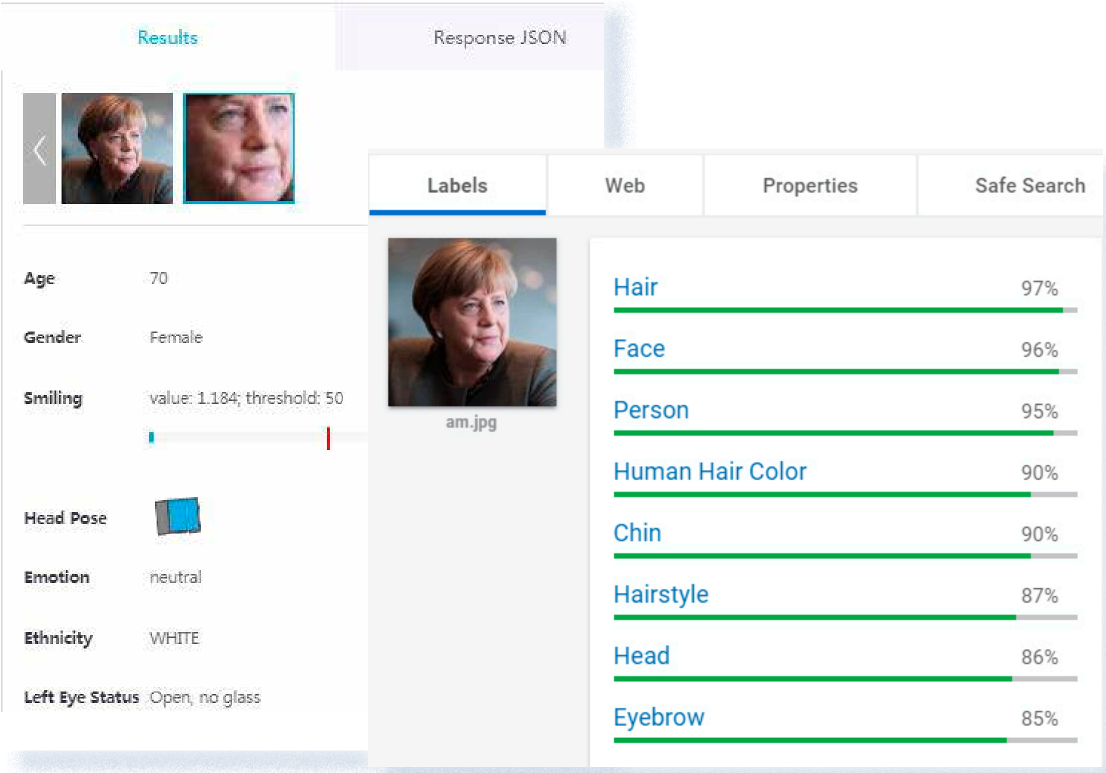


Figure 1

A machine-learning algorithm must first be trained on annotated images — data with defined class labels. Once an algorithm has been trained, testing it on data with no class labels can validate its performance. The input for the trained model in Figure 1 is an image of a cat; the output of the model is a class label. Note that the model can only recognise the classes it has been taught. If your goal is to recognise cats in images, your machine-learning algorithm must be trained using thousands of images of cats and thousands of images that do not contain cats. Images are usually complex and contain multiple objects, so comprehensive object recognition requires training using a very large number of annotated images containing and not containing all the different objects to be recognised.

Variations of object recognition exist. Two sub-categories are facial recognition and optical character recognition (OCR), each of which tailors the task of the model for a specific goal. In face recognition models are trained exclusively to recognise faces and, sometimes, the emotions they display. Whereas, OCR recognizes text and numeric



Age 70

Gender Female

Smiling value: 1.184; threshold: 50

Head Pose

Emotion neutral

Ethnicity WHITE

Left Eye Status Open, no glass

am.jpg

Hair	97%
Face	96%
Person	95%
Human Hair Color	90%
Chin	90%
Hairstyle	87%
Head	86%
Eyebrow	85%

Figure 2 On the left: analysis of Angela Merkel's image; on the right: Google Cloud Vision APIs results for the same image.

characters in an image. Digital images also hold information about the image file itself (metadata). This metadata is often referred to as EXIF data, and can include image size, location data, a smaller thumbnail of the image, and even the make and model of the camera used. EXIF data could be used as one of many steps in understanding the online information environment, although most social media platforms will strip an image of EXIF data in the process of sharing it.

There is no single technique that can provide a comprehensive description of what is depicted in any particular image. The different computer-based recognition capacities are further illustrated in Figure 2 below.

Google's Cloud Vision API — a commercial object recognition solution — can detect a person, face, head, and head-related features in an image. The Face++ API — a more extensive facial recognition solution — can recognise that the image depicts the face of a white female and estimate both her age and the emotion expressed.

2.1 Current Practices and Existing Tools

Many algorithms have now been trained and validated for object recognition tasks.[39] Their comparative performance regarding a variety of datasets has been well documented.[40] There are also several databases of annotated images that can be accessed to train and validate new models.[41] Bermeitinger et al. (2018) trained an algorithm to recognise Ukrainian political activist Stepan Bandera's face in conjunction with symbols of Russian or Ukrainian nationalism or fascism.[42] Their approach is a fine example of how machine-learning algorithms can be used for quantitative image analysis. The Bermeitinger approach solves a very specific problem, i.e. it looks for Bandera's face and then for symbols he is commonly associated with. The result is the classification of the message the image is trying to infer, e.g. Bandera is a Nazi. The downside of this approach is that it cannot be used for anything other than detecting Bandera's face and the set of symbols that also feature in the image search.

More general object recognition capability can be achieved by training an algorithm on some of the large annotated image datasets mentioned above. This task can also be outsourced to commercial object recognition services. These services lend their pre-trained models for various recognition tasks. Table 2 lists three commonly used commercial services that offer, among other things, object recognition.

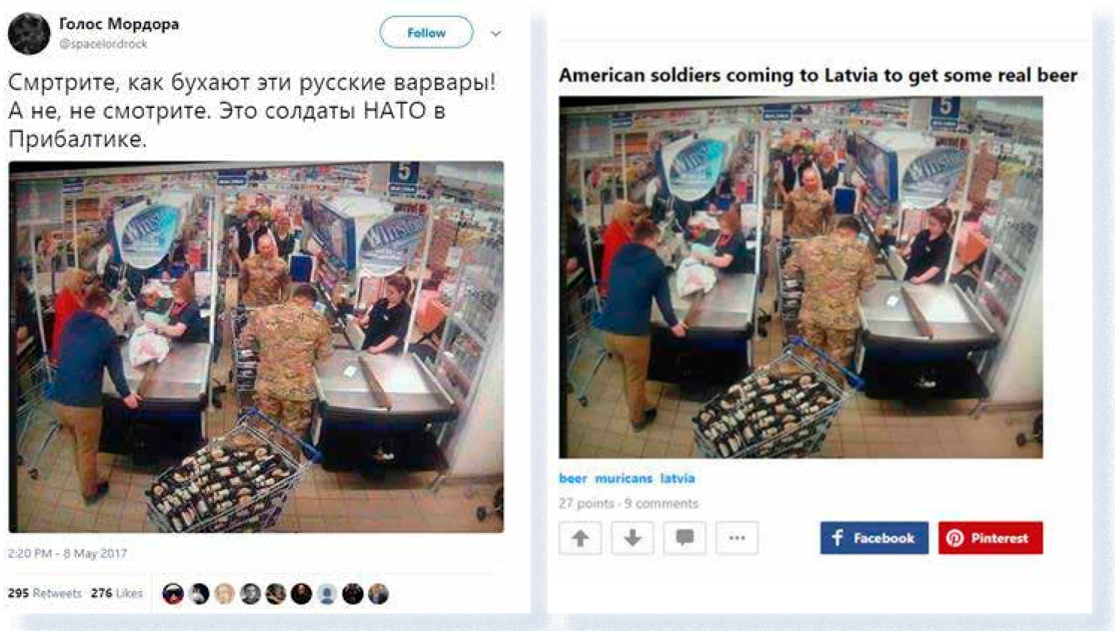
Solution	Description	Additional comments
Face++ API	**Facial recognition API endpoint**[43] Accepts an image or image URL and returns descriptions of faces found.	This programme has been suspended in the EU since the GDPR came into force.
Tensorflow Object Detection API	**Machine learning library** Provides a library of pre-trained object recognition models and annotated image data.	This service is a software library that requires the installation of TensorFlow and the downloading of the API codebase. This is a good platform to use when general object recognition falls short at executing specific tasks, e.g., recognizing logos in images.
Google Cloud Vision API	**Label API endpoint** Accepts an image or image URL and returns a list of objects found and their respective probabilities.	If text or numeric characters are found on the image, returns “text” or “number” labels, but not the actual characters.
	Document API endpoint Accepts an image or image URL and returns any text or numeric characters found.	
	Web entity API endpoint Accepts an image or image URL and returns named entities found in online texts associated with the image.	Searches for the named entities[44] found in online texts associated with a given image.

Table 2

3. CASE STUDY: NATO’s eFP

One of the goals for adding image recognition to the toolbox for analysis of the online information environment is to be able to detect mis- and disinformation communicated by means of images. The presence of foreign troops can be exploited by those engaged in adversarial information activities; incidents involving troops can be distorted, irrespective of whether they really took place. For example, in February

2017, an email was sent to a speaker of the Lithuanian parliament claiming that a group of German soldiers stationed in Lithuania had raped a 15-year-old girl.[45] It was deliberately false information shared to discredit the NATO troops in Lithuania. A smaller effort to discredit the troops stationed in Latvia took place in May 2017. We noticed an image trending across social media depicting American/NATO soldiers in Latvia buying a cart full of beer at a local chain store. See Figure 3 below.



Голос Мордора
@spacelordrock

Смртрите, как бухают эти русские варвары! А не, не смотрите. Это солдаты НАТО в Прибалтике.

2:20 PM - 8 May 2017

295 Retweets 276 Likes

American soldiers coming to Latvia to get some real beer

beer muricans latvia

27 points · 9 comments

Figure 3 On the left: image on 9gag; on the right: the same image on Twitter with the caption: 'Look, how much these Russian barbarians drink. Oh no, don't look. Those are NATO soldiers in the Baltics.'

The image appeared on Twitter, 9gag, and Facebook multiple times together with various derogatory messages. Upon further examination, it was discovered that the image predated the smear campaign by two years, and was initially published in the Estonian online space. Both cases demonstrate that NATO's eFP mission is a target for information activities, and there is a need for continuous monitoring. As of today, there is no known method for monitoring visual content of a particular topic; this chapter outlines a possible approach for monitoring and analysing the visual content regarding NATO's eFP mission.

3.1 Methodology

We collected data from Twitter, VK, and Instagram using Twitter Search API, VK API, and web scraping[46] during the period from 16 July 2017 to 10 January 2018. We used three search strings to capture both English- and Russian-language content regarding NATO activities in the Baltics and Poland, and the broader discussion on NATO in the Russian language. We then extracted the URLs of the images present in the identified posts and passed them through Google's Cloud Vision API label and web entities endpoints.

Further exploratory analysis of the resulting labels and web entities was conducted using Self-Organising Map (SOM) software developed by SOMTXT UG.[47] We used SOMs to display clusters of images similar to each other in terms of the web entities they returned. We then overlaid the resulting SOMs with heatmaps showing the frequency of these web entities in the dataset of the analysed images.[48]

3.2 Findings

In total we extracted 16,757 image URLs using the three search strings. The discussion about NATO in Russian yielded by far the greatest number of references to images — 14,264. When examining the volume of discussions regarding NATO in Russian over time, we observed a significant peak on 26 October 2017 (see Figure 4), the date of the NATO-Russia Council meeting. The most frequent labels extracted from underlying images for this date were *military,*

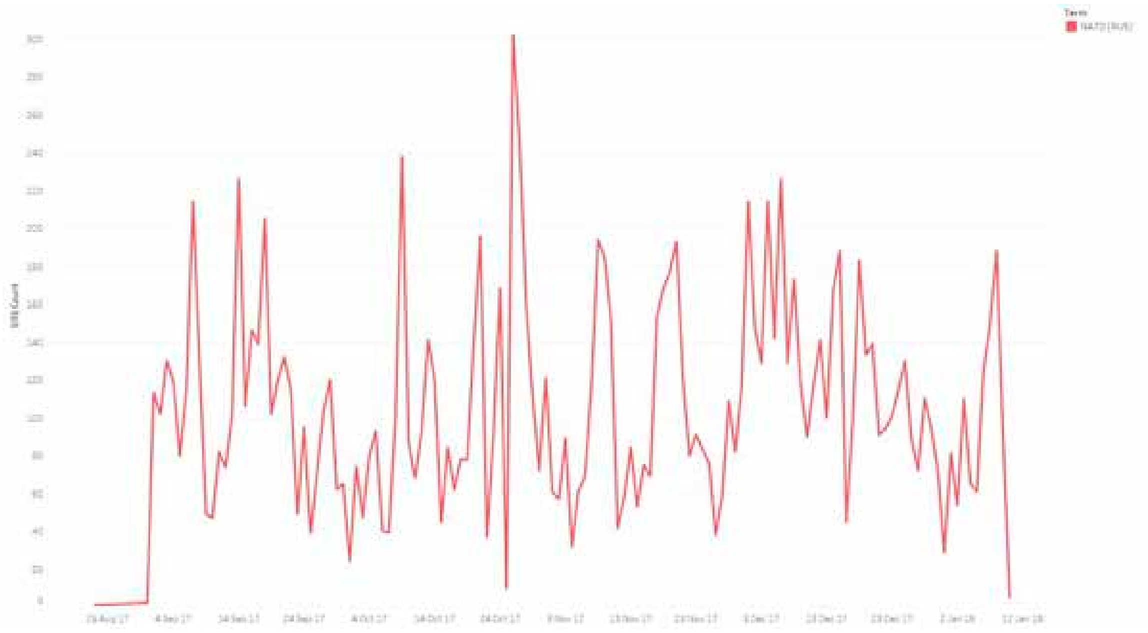


Figure 4 Top: Volume of images generated by discussions regarding NATO in Russian over time.

vehicle, public, professional, and *entrepreneur*. The web entities, however, were *NATO, Russia, military, Ukraine* and *exercise* (see Appendix 1). In this case the web entities were more specific and useful for understanding the visual echoes of the event.

The content on NATO in the Baltics and Poland showed a different pattern (see Figure 6). There was a significant increase in English language volumes on 1 August 2017, the date when an incident with Russian jets entering Baltic air space occurred. Smaller increases were observed shortly before and after the ZAPAD exercise on 10 and 23 September. [49]

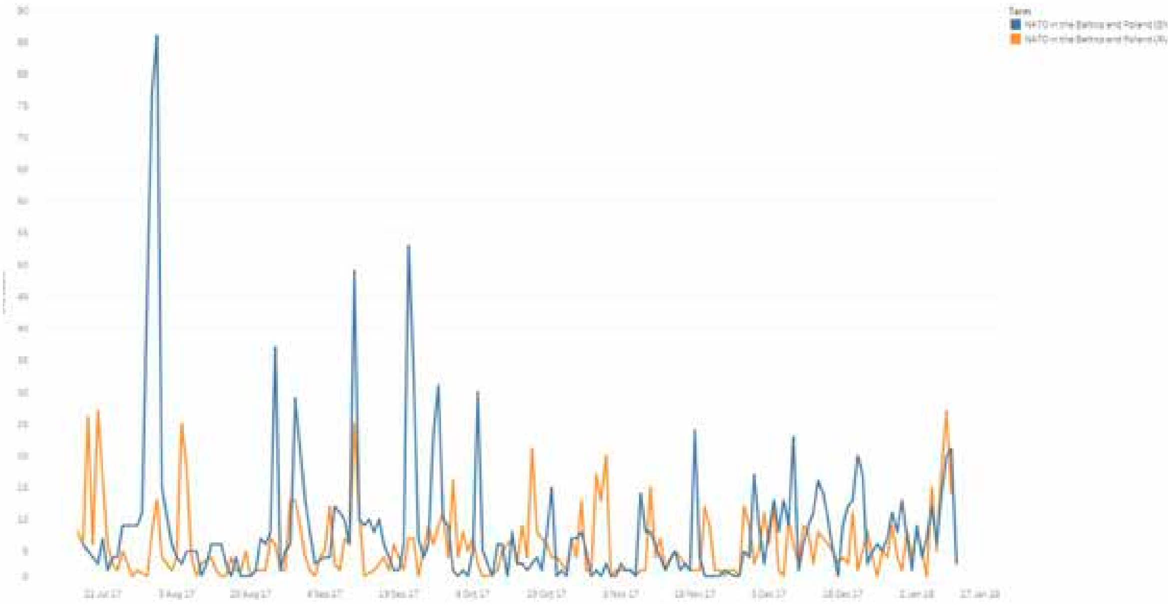


Figure 6 Volume of images generated by discussions regarding NATO in Poland and the Baltics in both Russian and English over time.

The most frequent web entities for images generated on 1 August were *aircraft, United States, Reuters,*[50] *Mikoyan MiG,*[51] *president, military, news, fighter, hornet.*[52] The respective labels were *aircraft, military, white, FA hornet, aerospace, airplane, engineering, fighter, force, jet*. In this case web entities were more descriptive of the underlying images, although to a lesser extent than in the previous example. Interestingly, the Cloud Vision API label endpoint succeeded in recognising a FA hornet, the fighter jet. This suggests Google's label endpoint is not only good in recognising generic objects, but also has the ability to recognise some very specific objects such as fighter jets.

The Russian-language content yielded the smallest number of references to images — 1,009 — and was not examined any further.

Self-organizing Maps (SOMs) and Visual Narratives

Exploratory analysis using SOMs was conducted with the generic 'NATO in the Russian language' dataset.[53] The range of topics was found to be very broad: from military equipment and media to international relations and past events. Terms related to the bombing of Yugoslavia were found to be popular. Terms related to Catalonia and the Ukraine crisis were found to be less frequent, but oftentimes used together. The most frequent words used in the NATO discussion in Russian language were *United States*, *Russia*, *Ukraine*, and *military*.

The underlying algorithm clustered terms according to how often they were used together, the more often the terms were used together, the more closely-clustered they appear. For example, the terms *Ukraine* and *war* were very close on the SOM (see Appendices) in the context of NATO. This means the words Ukraine and war are often used together and most likely comprise a separate topic. The SOM resulting from image labels (retrieved using Cloud Vision API labels endpoint) was too broad, and revealed no specific clusters in image labels. The map resulting from image web entities (retrieved using Cloud Vision API web entities endpoint) yielded much more meaningful results. To then identify the specific clusters, we first grouped six web entities that are near each other and are semantically connected, into one cluster. After this step we grouped 4–6 clusters, which are close to each other and are semantically connected, into one category. The result was 5 categories of 4–6 clusters (see Table 3). The web entities featured a number of military terms and many of the most frequent words were generic, therefore several semantically related terms could belong to more than one cluster.

Category	Clusters	Web entities (words)
Russia	Media, social media, Russian Federation, Russia's relations	Television, radio, YouTube, video, sputnik, power, Twitter, TV, journalist, regnum, newspaper, media, Moscow, Saint Petersburg, Duma, Ministry, Ukraine, Turkey, attack, China, Washington, United States, Egypt, Saudi Arabia

Category	Clusters	Web entities (words)
Armed forces and missile systems	Missile systems, ballistic missiles, naval forces, air forces	Weapon, rocket, system, intercontinental, range, ballistic, Kaliningrad, explosion, cruise, nuclear, missile, warfare, fleet, submarine, navy, sea, helicopter troops, landing, assault, amphibious dock, ship, air, aircraft, royal, carrier, sukhoi su, strike, fighter, airplane, Lockheed Martin, Boeing, raptor
Exercise and defence	Baltic air policing, infantry, eFP, ZAPAD, defence	Operation, Baltic, policing, vehicle, fighting, tank, battle, Shoygu, infantry, soviet, personnel, budget, Sweden, Bulgaria, Romania, Finland, Lithuania, Estonia, Poland, Latvia, strategy, exercise, Zapad, Belarus, Lavrov, Georgia, Minsk
NATO politics	Policy, Europe, organization, summit, Middle East	Defence, Donald Trump, enlargement, collective, peace, diplomacy, policy, Brussels, threat, Canada, Warsaw, powers, committee, Europe, headquarters, Atlantic treaty, resolute support, response, transformation, training, troop, summit, Afghanistan, action, joint, terrorism, Islamic Levant, Iraq, organisation, security, Israel, Iran, Syrian civil war
Current affairs and international relations	Turkish coup d'etat, war in Donbass, Catalonian referendum, Ukrainian politics, elections, Balkans	Ankara, Afrin, attempt, coup d'etat, Erdogan, assembly, relations, constitutional movement, declaration, independence, referendum, justice, development, accession, Crimea, parliament, protests, Donetsk, Lugansk, Luhansk, oblast, independent, Kiev, Ukrainian, union, Vladimir Putin, Donbass, Poroshenko Petro, solidarity bloc, corruption, Verkhovna Rada, prosecutor, statute, service, information, agency, presidential, election, diplomat, ambassador, embassy, republic, people, democratic, city, social, Macedonia, Skopje, national, Serbia, federal, Kosovo, Belgrade, Vučic Aleksandar, Serbian, province, autonomous, Yugoslavia bombing, Montenegro, foreign affairs, state

Table 3

As previously mentioned, the discussion regarding NATO in Russian language covered a wide range of topics from military equipment and media to international relations and past events. We overlaid the resulting SOM with a heatmap. The resulting heatmap shows the frequency with which terms were used — dark red indicates high frequency relative to the dataset and dark blue indicated low frequency relative to the dataset. If used in combination with a SOM, a heatmap can help reveal the visual narratives in the online discussion regarding NATO in the Russian language.

The **Current affairs and international relations** cluster featured several terms related to NATO bombing of Yugoslavia (Yugoslavian bombing, Kosovo, autonomous state).[54] When we overlaid the SOM with a heatmap that showed the frequency of web entities in the dataset, we found these terms to be quite frequent, which suggests this is a popular narrative in the Russian language space (see Figure 7) regarding NATO.

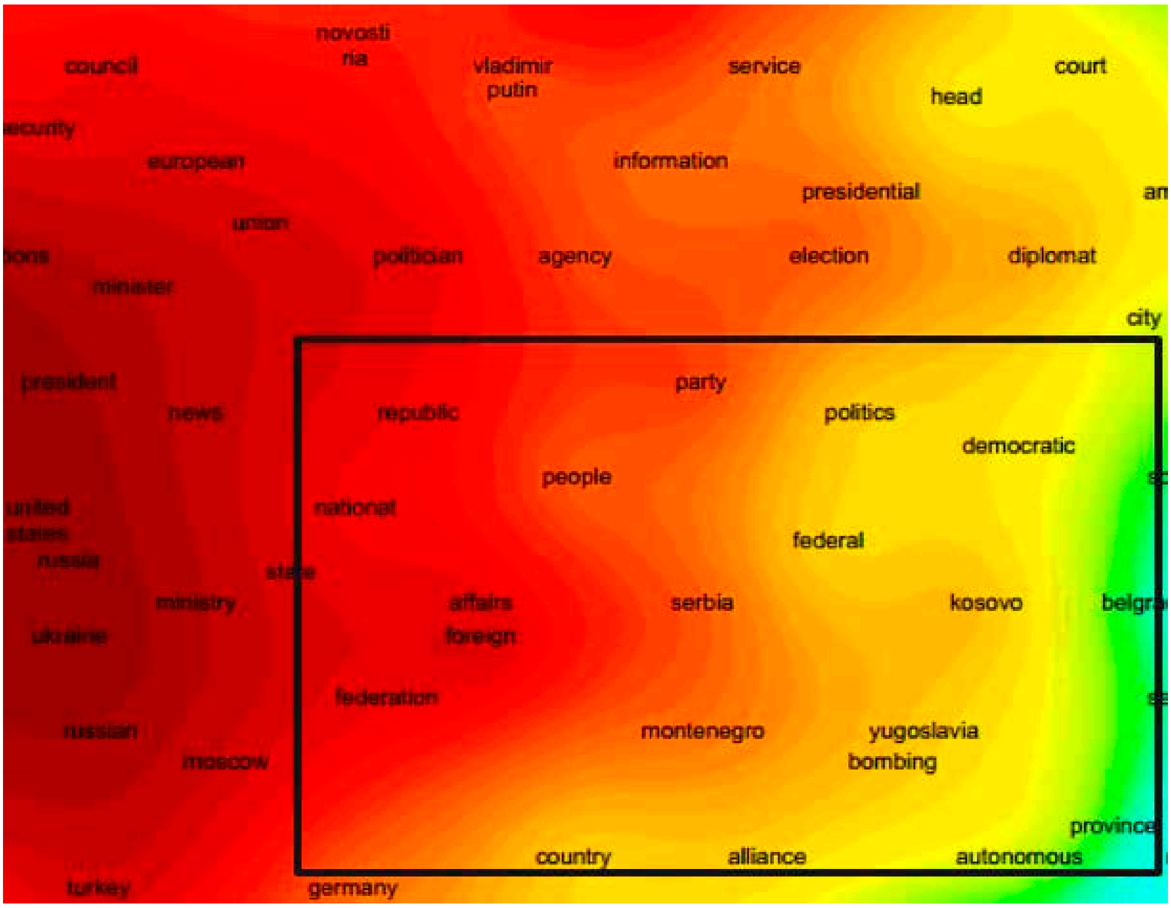


Figure 7 Part of Current affairs and international relations cluster overlaid with heatmap.

4. CONCLUSIONS

Fluctuations in the volume of images associated with a particular discussion were found to be a good indication of the offline events that drive online discussion. For example, on 1 August 2017, when three Russian jets entered the Baltic air space, there was an increase in image volumes associated with the online discussion of NATO eFP in the English language. From the various the image analysis methods we identified, the conclusions were as follows:

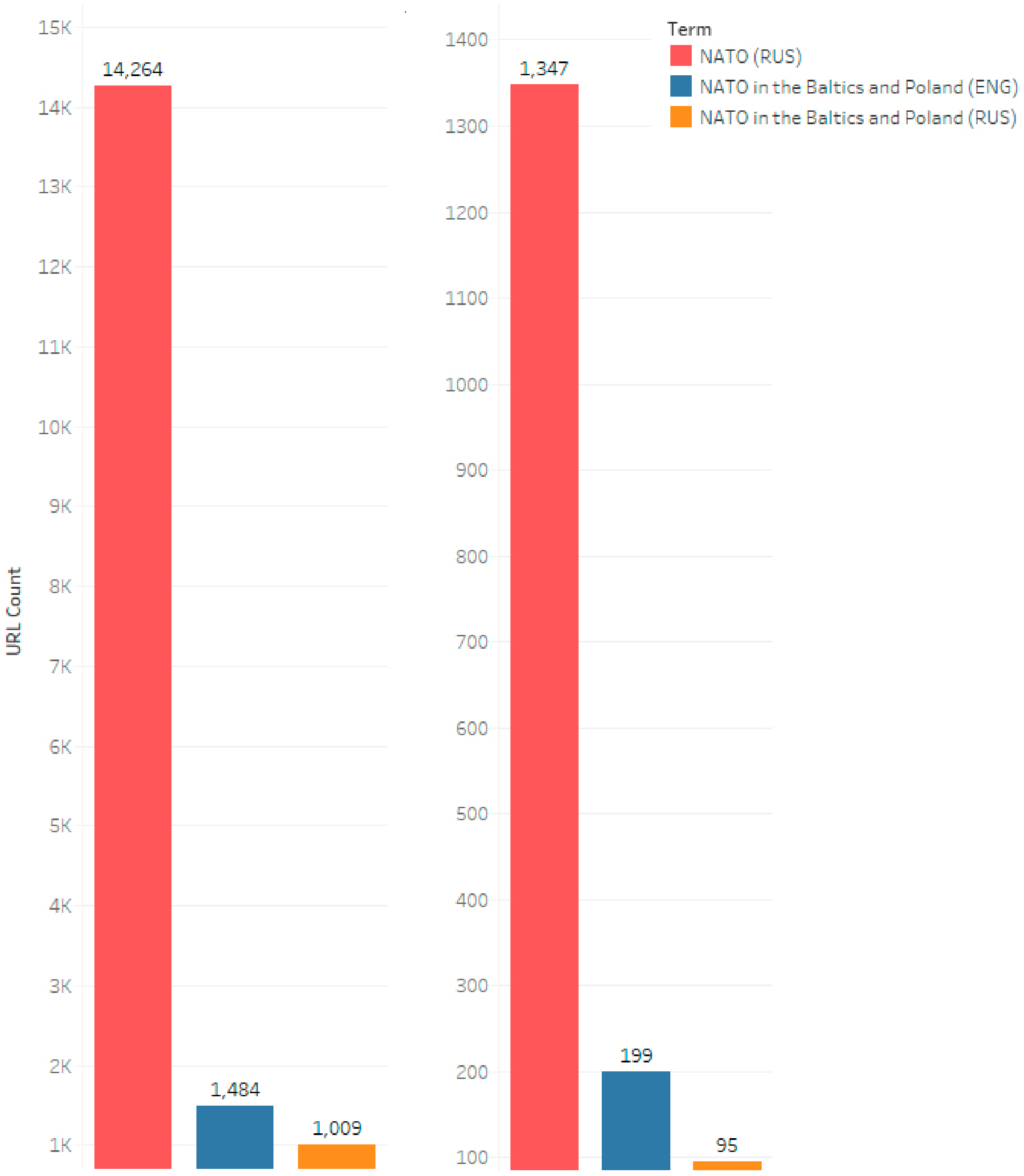
- Label recognition — yielded the least meaningful results;
- Web entity recognition — performed well at describing the gathered visual content.
- Label recognition paired with self-organising maps (SOMs) — yielded vague and non-descriptive maps;
- Web entities paired with SOMs — yielded descriptive maps that were further used to infer visual narratives in the gathered content.

We also found that the SOM algorithm used to cluster the web entities and reveal the underlying visual narratives performed well with a dataset of over 14,000 images. The other two datasets it was applied to (1,000–1,500 images) were found to be too small for the algorithm to yield meaningful results.

We conclude that using SOMs with web entities extracted from a dataset of unknown images is helpful for exploring and narrowing the field of topics depicted and discussed around images. The results, however, are greatly dependant on the volume of data provided to the SOM algorithm. A more in-depth examination of the images themselves is still necessary to reveal the underlying discussions. For the purpose of analysing the online information environment, generic object recognition solutions yield results that are too broad, and therefore offer no useful insights. An alternative is to examine the texts associated with the images. Further studies should be conducted comparing the visual and textual narratives of the same online content. It would be interesting to discover whether dissonance between the two can be an indication of sarcasm. As communicating via visuals becomes more popular, it is important to have this capability in place.

APPENDICES

On the left: total image URL count across search strings; on the right: count of images that contain texts



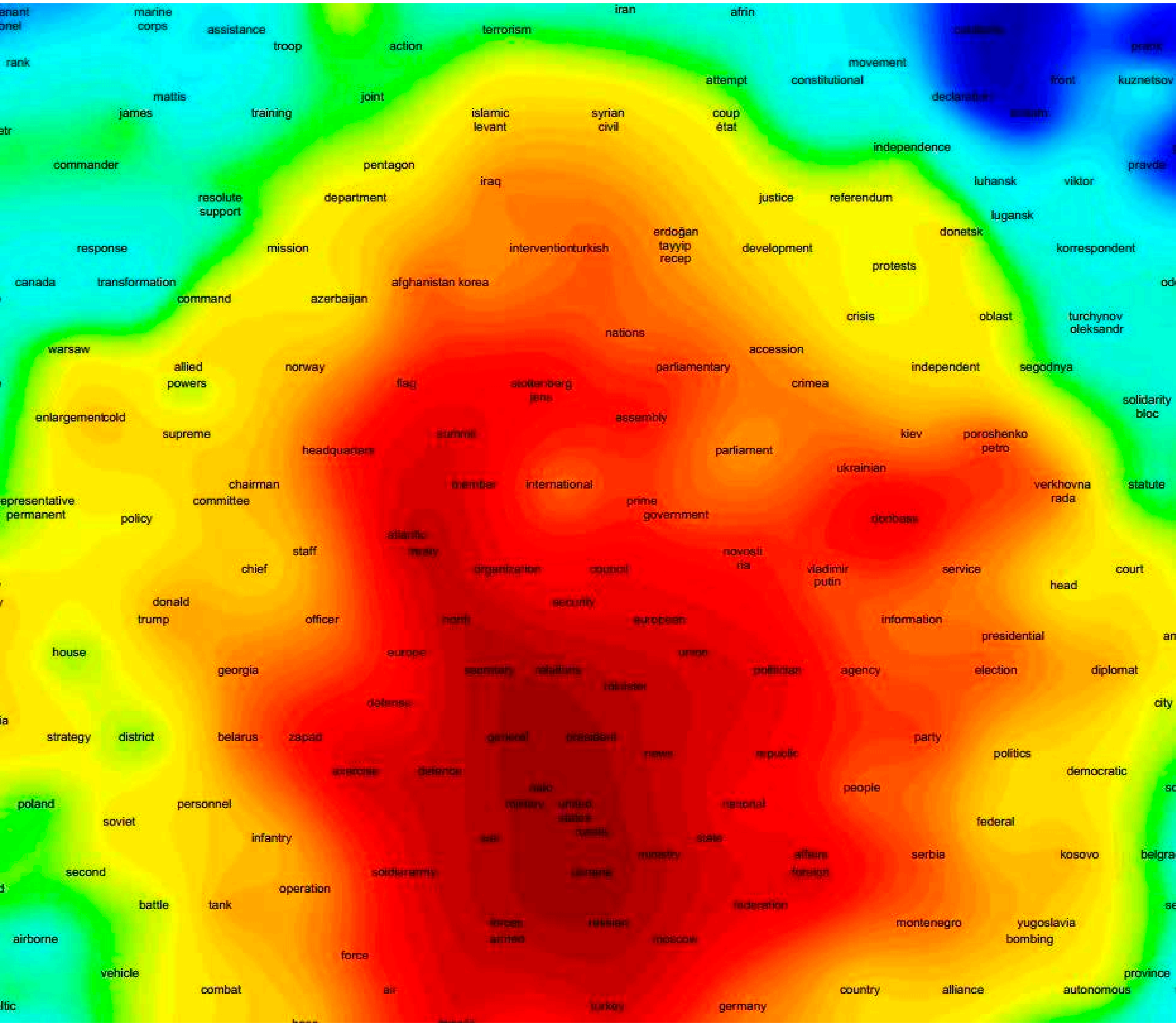
Most popular web entities and labels on 26 October 2017 (NATO in Russian language)

Web entities	Occurrences in text	Labels	Occurrences in text
NATO	**425**	*military*	**202**
Russia	**252**	*vehicle*	**120**
military	**177**	*public*	**85**
Ukraine	**114**	*professional*	**80**
exercise	**113**	*entrepreneur*	**74**
United States	**112**	*official*	**68**
general	**99**	*spokesperson*	**63**
relations	**93**	*speaking*	**60**
missile	**85**	*organisation*	**53**

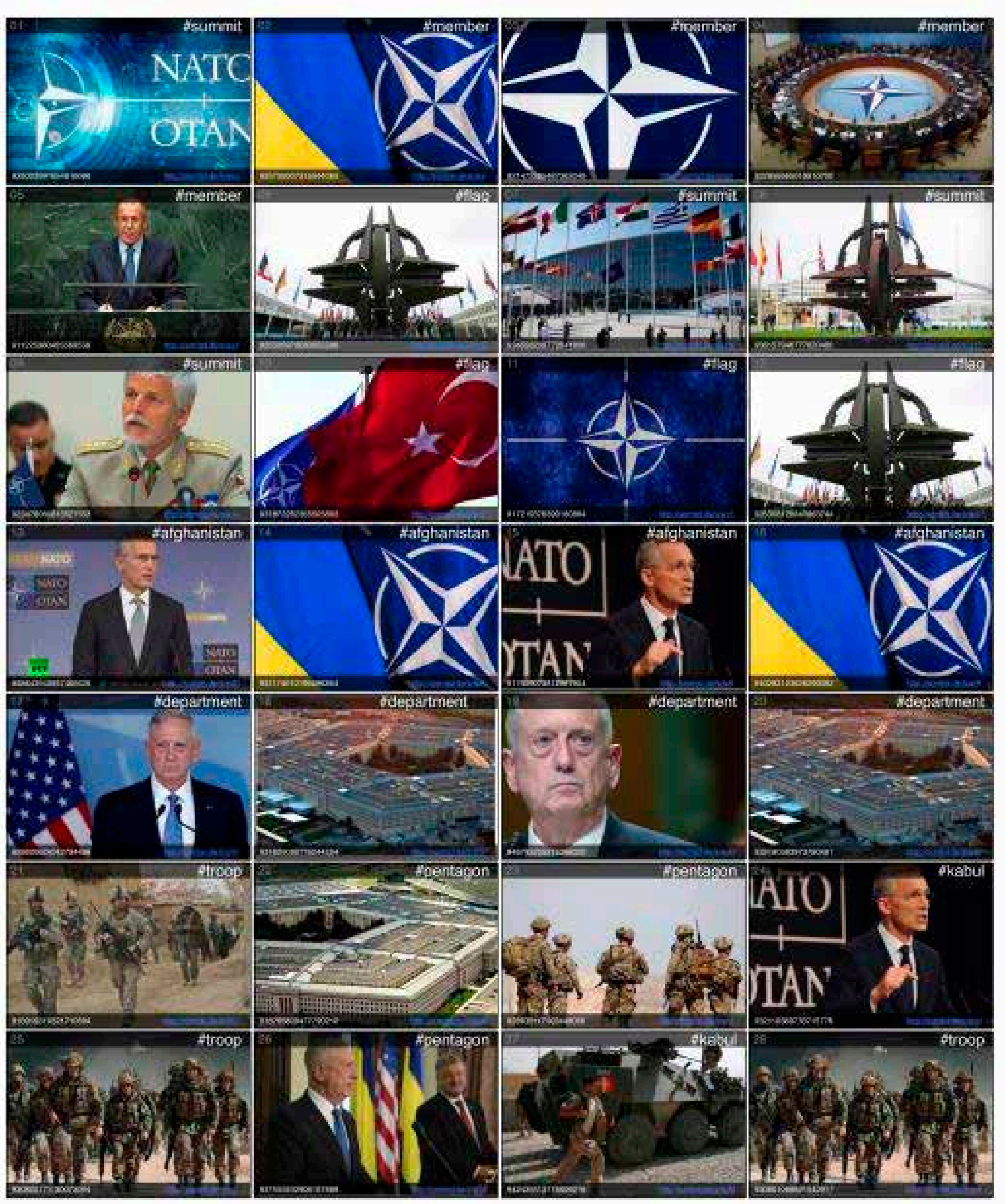
Most popular web entities and labels on 1st Aug, 2017 (NATO EFP in English language)

Web entities	Occurrences in text	Labels	Occurrences in text
aircraft	**82**	*aircraft*	**106**
United States	**73**	*military*	**52**
Reuters	**54**	*white*	**37**
Mikoyan MiG	**42**	*FA hornet*	**26**
president	**38**	*aerospace*	**25**
military	**28**	*airplane*	**25**
news	**26**	*engineering*	**25**
fighter	**22**	*fighter*	**25**
hornet	**22**	*force*	**25**

Generic NATO in Russian Language SOM with heatmap



Images in the ‘Summit’ cluster



REFERENCES

1. Bermeitinger, Bernhard & Radisch, Erik & Howanitz, Gernot. (2018). Contextualizing Bandera: Ein Distant Watching-Ansatz, https://www.researchgate.net/publication/323507402_Contextualizing_Bandera_Ein_Distant_Watching-Ansatz
2. DAESH Information Campaign and its Influence, NATO Strategic Communications Centre of Excellence, Riga, 2016, https://www.stratcomcoe.org/download/file/fid/5806
3. "Lithuania Looking For Source of False Accusation of Rape by German..." 2018. Reuters, *U.S.* https://www.reuters.com/article/us-lithuania-nato/lithuania-looking-for-source-of-false-accusation-of-rape-by-german-troops-idUSKBN15W1JO
4. "Mary Meeker'S 2016 Internet Trends Report: All The Slides, Plus Highlights". 2018. Quartz. https://qz.com/697050/mary-meekers-2016-internet-trends-report-all-the-slides-plus-highlights/
5. "Review Of Deep Learning Algorithms For Object Detection". 2018. Medium. https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852.

04

THE EFFECTS OF BANNING THE SOCIAL NETWORK VK IN UKRAINE

Anton Dek, Kateryna Kononova, Tetiana Marchenko

ABSTRACT

In May 2017, the President of Ukraine put into effect a decision of the National Security and Defence Council (NSDC) to impose economic sanctions on 468 Russian companies. The largest Russian social network, VKontakte (VK), was banned, among others. Because of the ban, the audience for VK in Ukraine decreased significantly and the social network site dropped out of the top ten most popular sites in the country.

To better understand the effects of the VK ban, a study was organised to monitor changes in posting dynamics, to analyse user demographics before and after the ban, and to identify the rhetoric used in posts before and after the ban. The dataset includes more than 300,000 Ukrainian VK user profiles. Because the ban does not apply to the territories occupied by Russia, the study examined two target regions: government-controlled areas (GCA) subject to the ban, and non-government-controlled areas (NGCA) where the ban was not imposed. The study took place between 1 May 2016 and 14 June 2018. This period was divided into three intervals: before the ban, the first 'user exodus', and the second 'user exodus'.

Our analysis shows that VK is markedly less popular in the area controlled by the Ukrainian government (19% less, compared to NGCA). However, those few users left after the ban are more active, producing 4.37 times as much content as those in non-government-controlled-areas. Moreover, these VK profiles are more densely connected.

To study the rhetoric appearing in user posts, we used a clustering algorithm that could identify accounts posting about ideological issues. When compared with the majority of other profiles, the characteristics of ideological posts stood out. Before the ban a typical VK-user would write, on average, one post every four days; after the ban the frequency dropped to one post every ten days. Ideological users[55] were notably more active — they wrote four posts per day before the ban and 1.6 posts after the ban. Ideological users were also significantly more connected — after the ban the average number of friends for an ideological user grew from 197 to 501, and such users subscribed to 2.25 times more groups than typical VK-users.

Our analysis of reposts from ideological groups showed that although the number of users decreased by a factor of three, the activity level of these groups remained unchanged. However, taking into account that most pro-Ukrainian groups left VK after the ban, those who continued to use the network were increasingly posting to an echo-chamber.

We conclude that the VK ban was effective in some ways. The network lost a significant portion of its audience and its popularity

decreased, but those users who retained their VK profiles after the ban became more connected on average and began consuming more information from a greater number of groups. Despite the fact that the number of ideological users also dropped, those who continued to use VK lost their opposition, so rhetoric became less diverse.

INTRODUCTION

On 28 April 2017, the National Security and Defence Council (NSDC) adopted a decision to institute economic sanctions against 468 Russian companies. President Petro Poroshenko put the decision of the NSDC into effect on 15 May, 2017 in accordance with Ukrainian government Decree No. 133/2017.

The Ukrainian Law 'On Sanctions' states that sanctions have been imposed in order 'to protect the national security and territorial integrity of Ukraine, to counteract terrorism, and to prevent violations of the rights, freedoms, and interests of the citizens, society, and state of Ukraine'.

The largest Internet companies banned included the Russian social networks Vkontakte (VK) and Odnoklassniki, the search engine company Yandex (including its sites using the .ua domain), and the email service Mail.ru.

According to SimilarWeb,[56] the audience for VK in Ukraine decreased by more than 60% as a result of the ban, but according to Google Trends the decrease in users reached almost 80%.

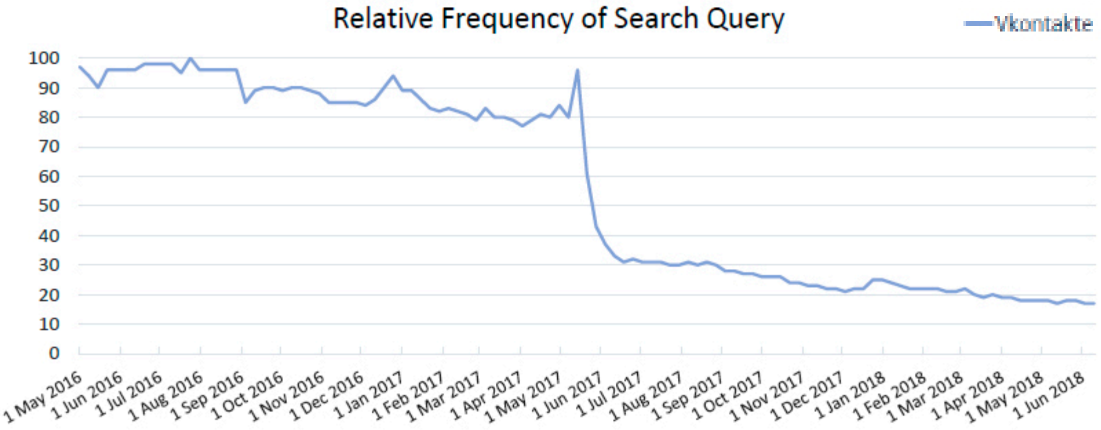
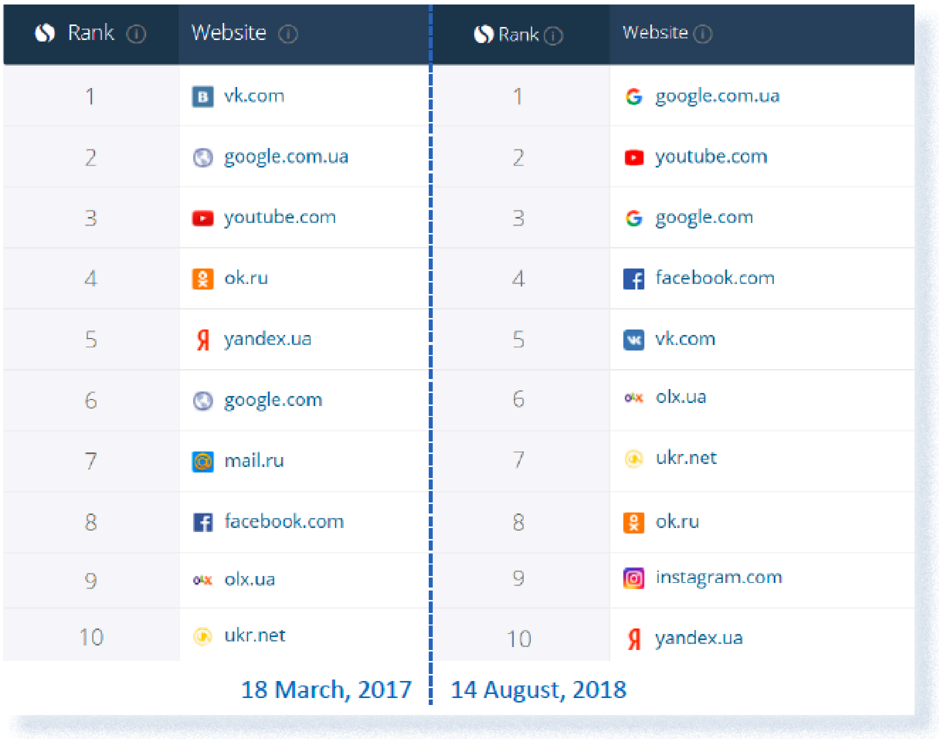


Figure 1: Relative Frequency of Searching by "VKontakte" Keyword in Ukraine[57]

According to Top Sites Ranking for All Categories in Ukraine, VK was Ukraine's most visited website before the ban.[58] After the ban, in the second half of 2017, VK quickly dropped out of the top five, and then from the top ten most-visited sites by Ukrainians. However, in 2018, the online social network has returned to its previous popularity. According to SimilarWeb[59] and Alexa Rank[60], VK is now among the top five most-visited sites in Ukraine.



Rank	Website	Rank	Website
1	vk.com	1	google.com.ua
2	google.com.ua	2	youtube.com
3	youtube.com	3	google.com
4	ok.ru	4	facebook.com
5	yandex.ua	5	vk.com
6	google.com	6	olx.ua
7	mail.ru	7	ukr.net
8	facebook.com	8	ok.ru
9	olx.ua	9	instagram.com
10	ukr.net	10	yandex.ua
	18 March, 2017		14 August, 2018

Figure 2: Top Sites Ranking for All Categories in Ukraine

Who are the Ukrainian VK users today and what topics do they discuss? Was the ban effective and how did it affect national security?

To answer these questions, researchers carried out the following tasks:

1. Posting dynamics over time
   - investigated traffic dynamics
   - identified changes in posting trends

2. Demographics before and after the ban
   - analysed gender and age distribution of users in each period
   - analysed the number of posts, friends, and group distribution for each user
   - investigated the characteristics of the user exodus

3. Posts topics and rhetoric analysis

- identified and analysed the dynamic of clusters of topics discussed for each period (including the ratio of pro/anti-Ukrainian rhetoric)
- analysed the dynamics of comments on ideologically-tinged posts
- analysed the activity of ideological groups

The study covers the period from 1 May 2016 to 14 June 2018. To analyse the results of the VK ban, we collected a sample of 870,174 Ukrainian VK users. Of these, 98% were not blocked or deleted;[61] and 0.7% of the profiles changed or obscured their location.[62] As only active users were included in the final sample, the dataset consists of 315,697 profiles.[63] Information about these profiles was downloaded to the project database.

The study examined two target regions:

- Territories under Ukraine's control, or government-controlled areas (GCA), where the ban on VK was imposed.[64]
- Territories occupied by Russia, including Crimea[65] and parts of the Donetsk and Luhansk regions, or non-government-controlled areas (NGCA), where sanctions could not be imposed.[66]

The total area of the occupied territories is 47,000 km^2, where 13.6% of Ukrainians reside.

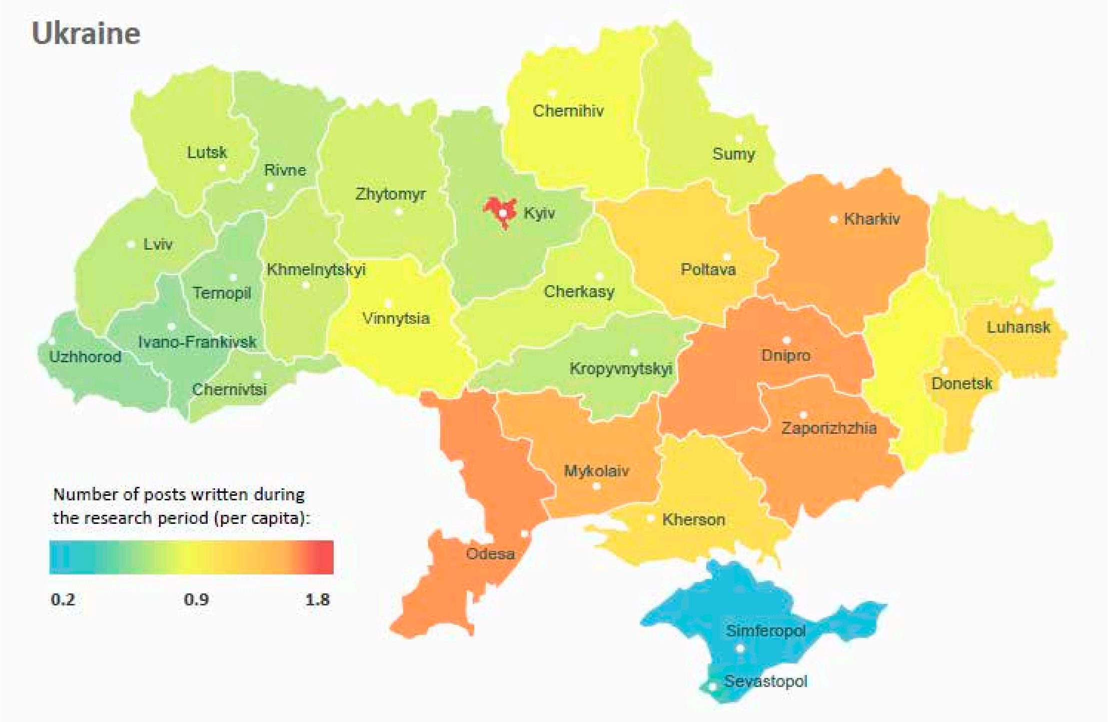


Figure 3: Number of Posts Written During the Research Period (Per Capita)

1. POSTING DYNAMICS OVER TIME

Analysing the dynamics of users' posts in the sample, it should be noted that before the ban the total number reached 101,000[67] per day. After the ban, the mean number of posts decreased by 53% to 38,000[68] on average. Since April 2018, the traffic decreased by 10,000 again and stabilized at that level.[69] Based on these observations, we introduced the following periodization of the study:

- before the ban, from 01/05/2016 to 04/06/2017[70]
- during the first user exodus, from 05/06/2017 to 08/04/2018
- during the second user exodus, from 09/04/2018 to 14/06/2018

The first user exodus could be explained by law-abiding citizens' rejection of VK, and by the inability of the rest to use virtual private networks [VPNs]. The second wave may have been caused by those who remained realising they had lost their audiences, making activity on VK less interesting to them. Both waves of users leaving VK were accompanied by peaks of activity on other social media. There was a significant surge in the use of Facebook search queries in

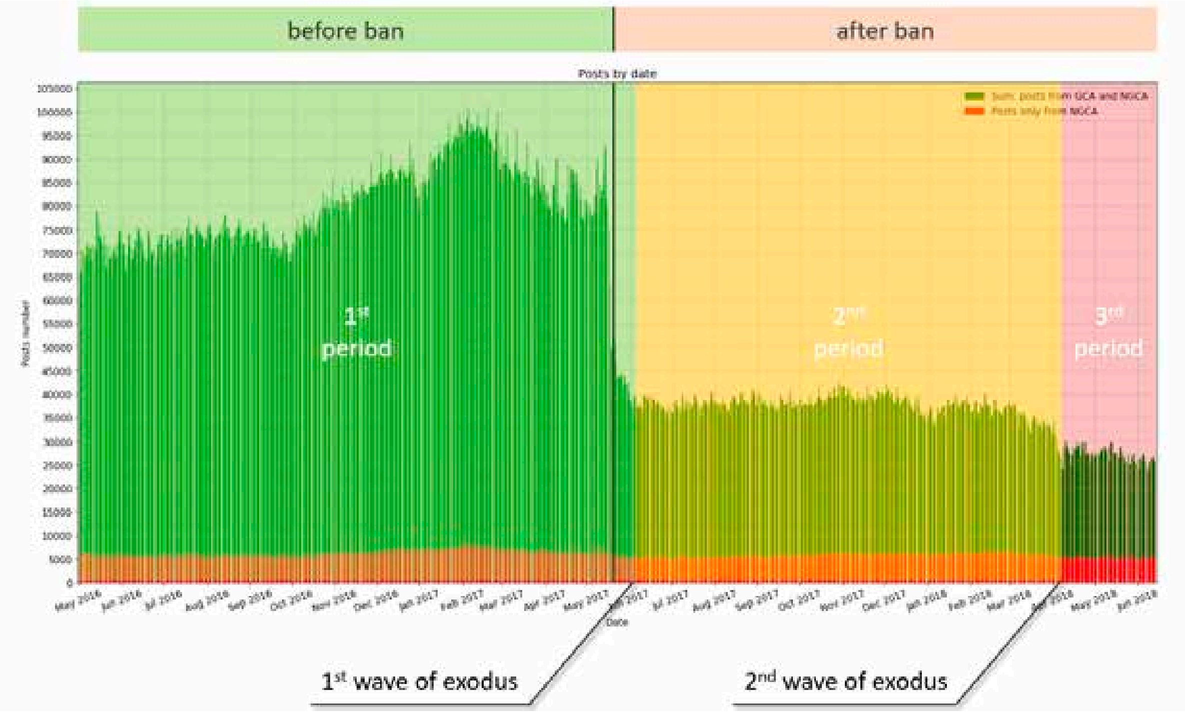


Figure 4: Posting Dynamics During the Research Period

mid-May 2017, and two spikes in the use of Telegram search queries in May 2017 and April 2018.

Nevertheless, while comparing the share of VK users in occupied and government-controlled areas, please note that despite the significant decrease in activity, VK dissemination in territories under the ban remained 28% higher than in the NGCA; the total number of posts was 5.5 times higher in those territories (NGCA).[71] In addition, user activity became more similar throughout Ukraine after the ban, although before the ban it had been almost two times higher in the GCA.[72]

2. DEMOGRAPHICS BEFORE AND AFTER THE BAN

While distribution according to sex was similar in the GCA and the occupied areas before the ban (51.6% female vs 48.4% male), after the ban the share of female profiles increased to 55% in both territories.

As shown below, the average age of users (as stated on their profiles) did not change in the occupied areas, while the average age of users in the GCA decreased.[73] This may have been a result of younger people using mobile phones to browse the internet and access the social network, [74] compounded by the fact that older people may be less familiar with VPNs, which could help them get around the ban.

Our analysis of the distribution of posts on VK showed that users from the GCA were significantly more active before the ban, but after the ban the number of users who wrote fewer than 20 posts increased by 23%. The number of those who posted a minimum of once every two days decreased by almost 10%. We thus conclude that, as a result of the ban, not only did the VK audience shrink, but the posting frequency of those who continued to use the network was reduced. We consider these to be positive effects of the ban. User activity in occupied territories did not change significantly during the period studied.

Our analysis of the distribution of users' friends demonstrated that less-connected users were more likely to leave the social network. The share of those who had fewer than 25 friends decreased by ~7% in GCA; in comparison, the share of those who had fewer than 50 friends decreased only by ~2.3%. The share of those who had 75 or more friends increased slightly. Thus we conclude that the average user who

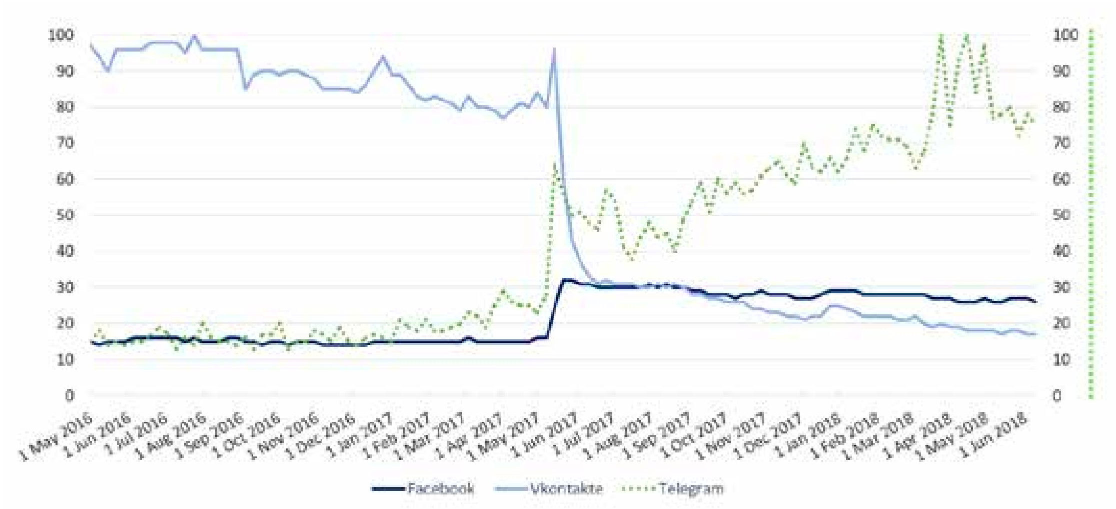


Figure 5: Relative Frequency of Search Queries Related to Vkontakte, Facebook, and Telegram in Ukraine[75]

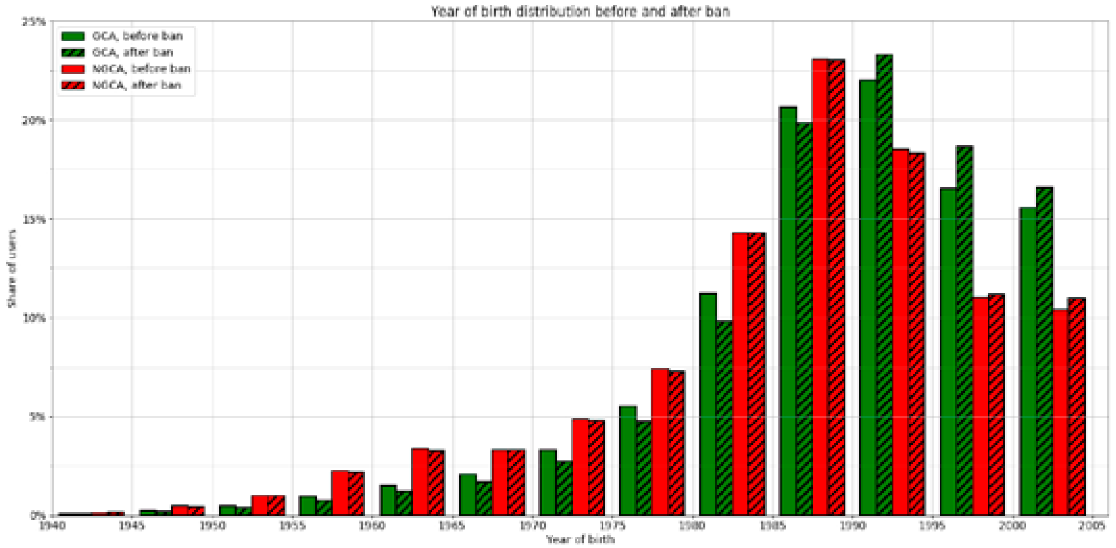


Figure 6: Age Distribution Before and After Ban

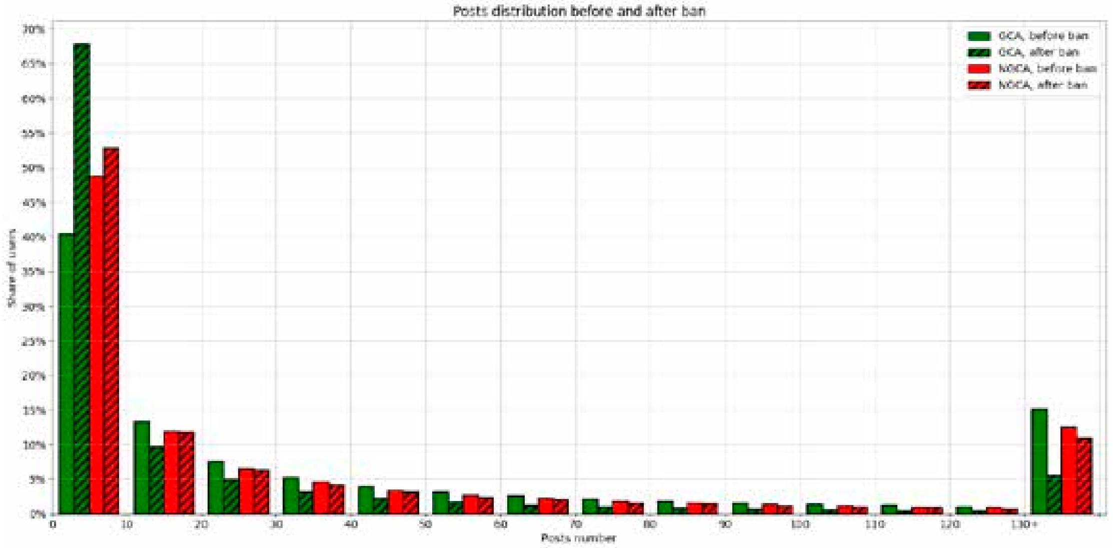


Figure 7: Distribution of Posts on VK Before and After Ban

continued to use VK after the ban was more connected. Users in the occupied territories also experienced an increase in the number of friends they had on VK, but this is likely to have been the result of typical network development.

Our analysis of the distribution of users' subscriptions to various groups shows that the number of subscriptions before the ban did not differ significantly in the two territories. After the ban, the number of groups followed increased in both regions, however the subscription rate was higher in the GCA. The number of users who subscribed to fewer than twenty groups decreased by 8%, and the share of those who follow 300 or more groups increased by 2%. In general, after the ban, all users began

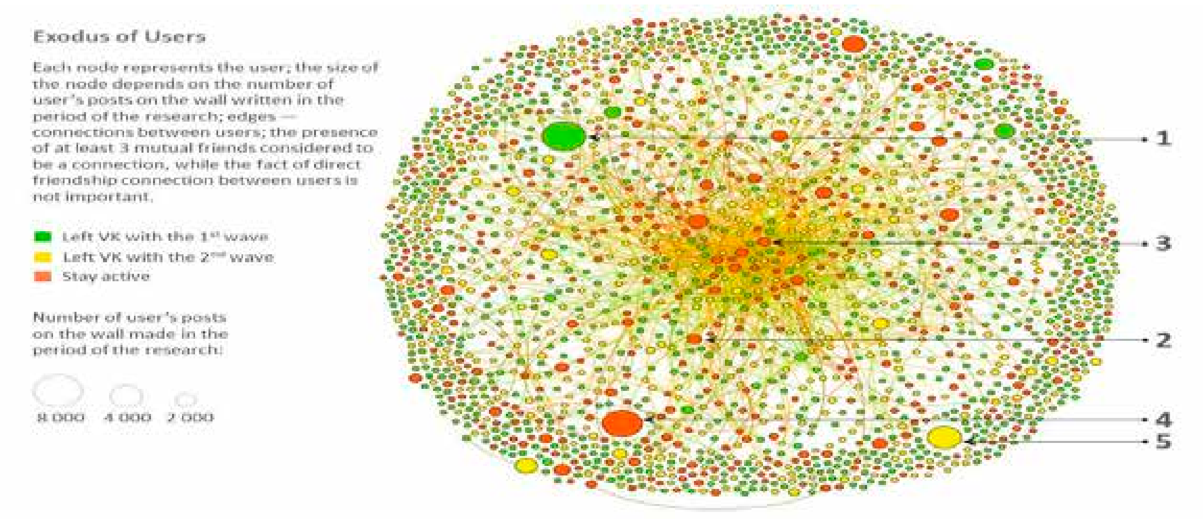


Figure 10: Connections Between a Sample of 2000 Random Users from the GCA

Details regarding the labelled representative nodes:

1. The most active user account in the 1st period generated 7,100 posts (or 17.7 posts per day). This user left VK with the first exodus and generated no posts in either the 2nd or 3rd periods.[77] This user's content was non-ideological; mostly consisting of reposts from other groups. This profile was located in Eastern Ukraine.
2. The most active user account in the 2nd period generated 1,399 posts (or 4.6 posts per day). This user was also quite active in the 3rd period generating a total of 108 posts (or 1.6 posts per day), but posted only five times in the 1st period. The content was non-ideological. The profile was located in Western Ukraine.
3. The most active user account in the 3rd period generated 444 posts (or 6.7 posts per day). The user sold sport shoes. All the posts contained photographs of shoes. The profile was located in Western Ukraine.
4. This account belongs to one of the most active users for all three periods, however, after the ban the posting frequency decreased drastically from 6,168 posts (or 15.3 posts per day) to 238 posts (or 1 post per 1.5 days). The profile was located in Central Ukraine. The user's content was pro-Ukrainian.
5. This user left VK with the second exodus. Before the ban this user generated 5, 200 posts (or 13 posts per day), but only a total of 31 posts (1 post per 10 days) after the ban. The profile was located in Central Ukraine. The user's content was non-ideological, mostly consisting of material reposted from the groups the user subscribed to, accompanied by the author's comments.

consuming more information from the groups they were following, but users from the GCA did this to a greater degree.

A social graph (see Figure 10) depicting the connections between ~ 2 000 random users from the GCA was created to analyse the user exodus in greater detail. Only active users were included in the sample.[76]

As shown in the diagram, the most-connected users, located in the central part of the graph, mostly continued to use VK (marked red), while less-connected users, located on the periphery, left the social network during either the first exodus (marked green) or the second exodus (marked yellow). In addition, users who posted more frequently were also more likely to stay (the majority of large nodes are red).

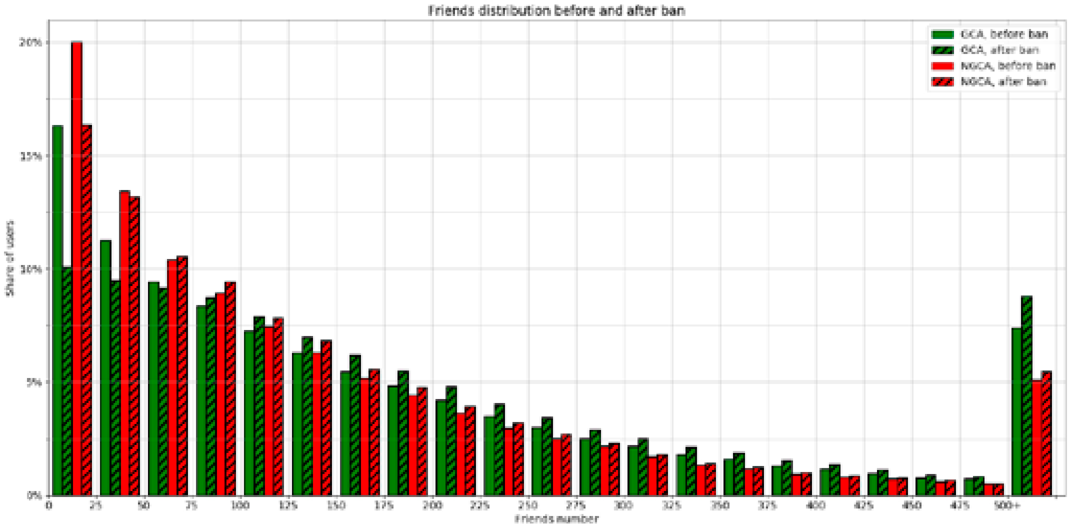


Figure 8: Distribution of Number of Friends per User Before and After Ban

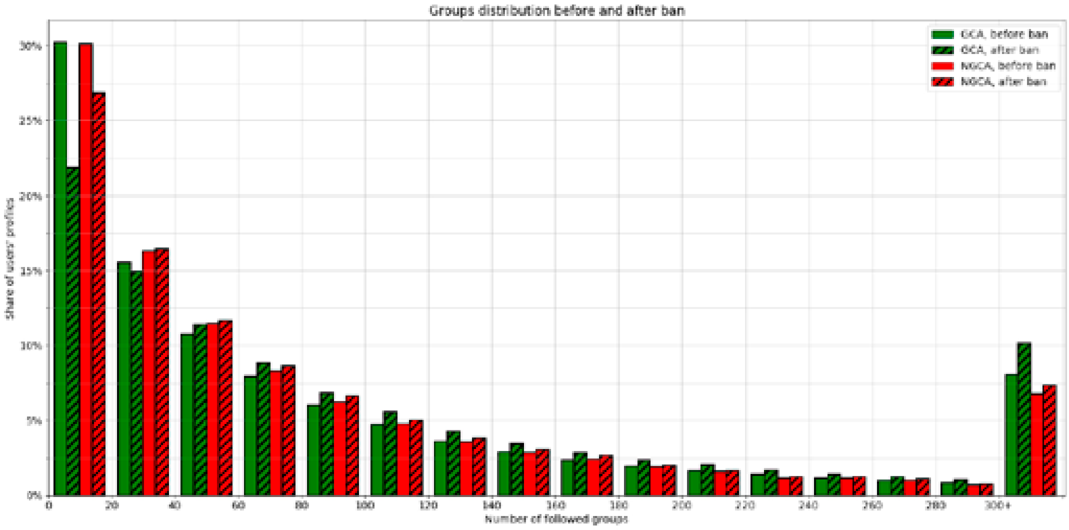


Figure 9: Distribution of Number of Groups Followed per User Before and After Ban

The distribution of posts in the sample shows that user activity decreased significantly after the ban. In the 1st period (01/05/2016 to 04/06/2017) only a third of the users wrote fewer than 10 posts; in the 2nd period (05/06/2017 to 08/04/2018) their share reached 74%; in the 3rd period (09/04/2018 to 14/06/2018) almost 95% of users became low-frequency posters. The share of those who generated 130 or more posts decreased by a factor of three after the ban, and in the 3rd period such active users were almost gone.

After the ban, VK users became more connected. The share of those who had fewer than 25 friends decreased by 15% during the 2nd period. In the third period it decreased threefold again; only 4% of users who had fewer than 25 friends continued to use the social network.

In addition, during all three periods, the most active users tended to consume more and more information from groups.

3. POST TOPICS AND RHETORIC ANALYSIS

To identify ideologically-tinged traffic, researchers created two random samples of about 300,000 posts for before and after the ban.[78] The posts sampled were written in Russian or Ukrainian only, were longer than 140 characters, and were not marked as spam.[79]

In these samples, the share of posts written by users from the GCA decreased by 11 percentage points after the ban,from 91% to 80%. All posts were written or reposted by the owners of the walls; the number of unique users was 75,089 in the first set and 48,766 in the second set. Among these, 91.6% of profiles were located in the GCA before the ban, but only 84.6% remained after the ban.

First, clusters of ideological posts were identified in each sample.[80] Second, posts from these clusters were vectorised and clustered again to refine the data. Once this was accomplished, we were able to identify ideological clusters in the first sample and nine in the second sample.

The number of ideological posts after the ban increased by 1.22 times compared with the period before the ban.[81] Five clusters are present in both samples (See Table 1), while the others were unique for the selected period. According to the results of the clustering process, the following significant topics were not addressed in the second period: 'KrymNash', 'Patrol Police', 'Pro-Ukrainian rhetoric', but a new topic,

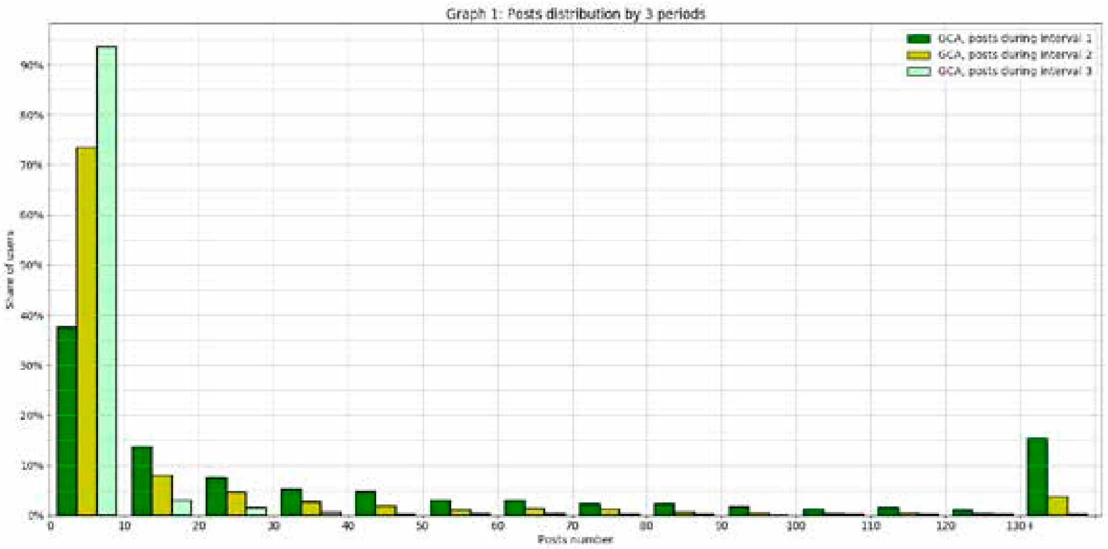


Figure 11: Distribution of Posts Written by Random Users During All Three Periods Studied.

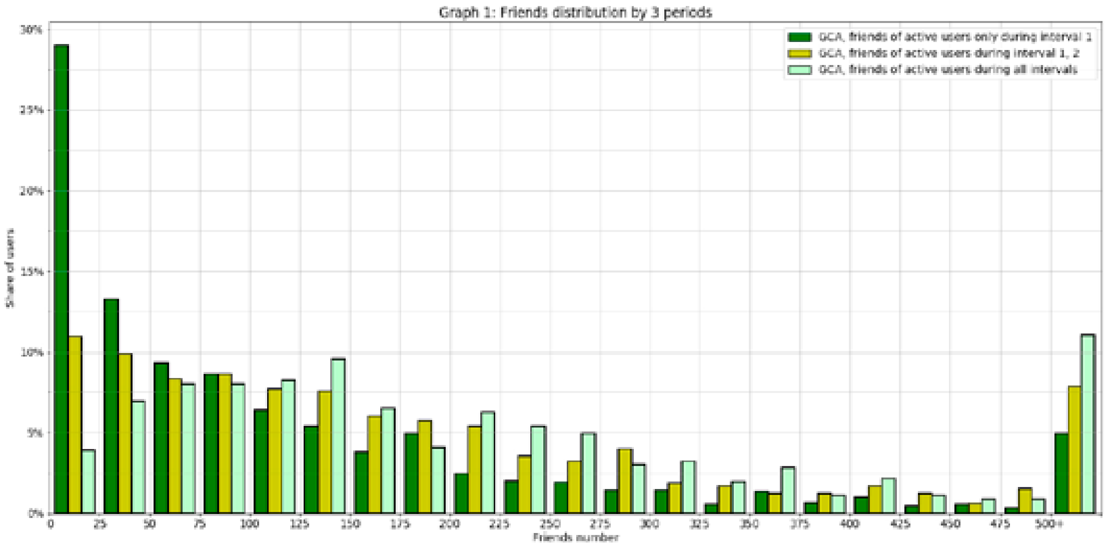


Figure 12: Distribution of the friends of random users during all three periods studied.

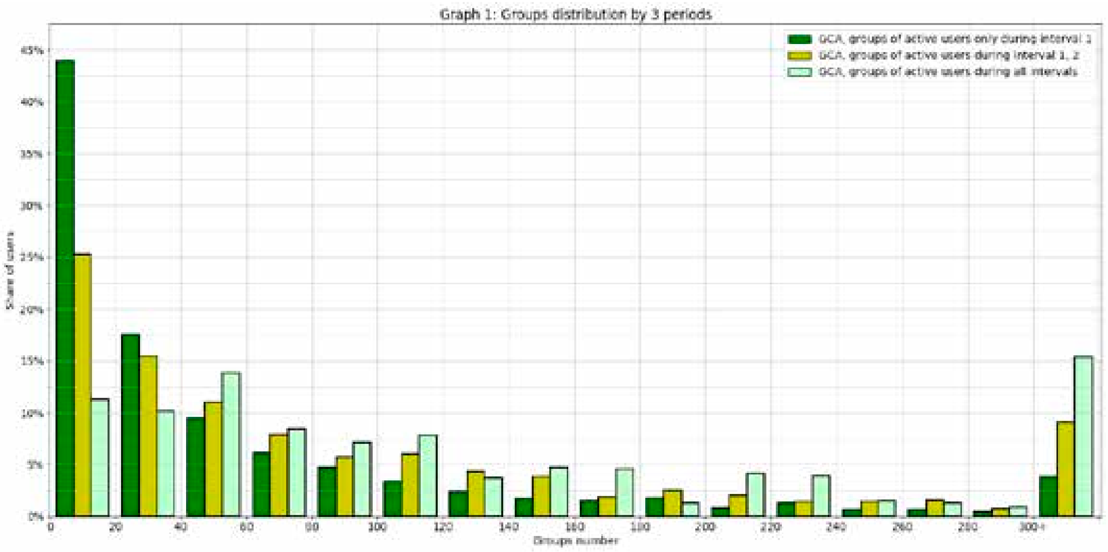


Figure 13: Random Users' Groups Distribution During Three Periods

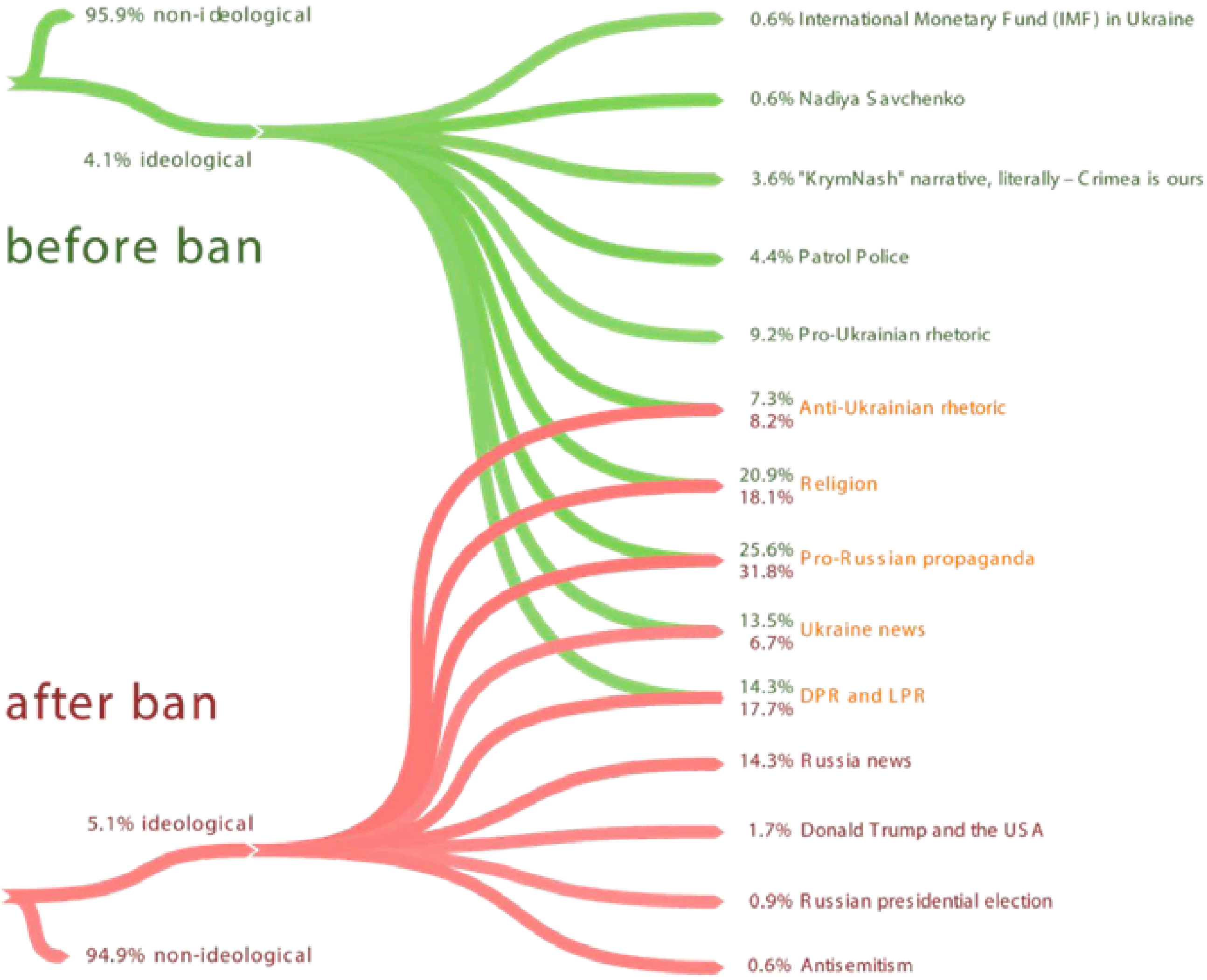

95.9% non-ideological
4.1% ideological
before ban
after ban
5.1% ideological
94.9% non-ideological
0.6% International Monetary Fund (IMF) in Ukraine
0.6% Nadiya Savchenko
3.6% "KrymNash" narrative, literally – Crimea is ours
4.4% Patrol Police
9.2% Pro-Ukrainian rhetoric
7.3% 8.2% Anti-Ukrainian rhetoric
20.9% 18.1% Religion
25.6% 31.8% Pro-Russian propaganda
13.5% 6.7% Ukraine news
14.3% 17.7% DPR and LPR
14.3% Russia news
1.7% Donald Trump and the USA
0.9% Russian presidential election
0.6% Antisemitism

Figure 14: Cluster Dendrogram

'Russian news', appeared. 'Pro-Russian propaganda' notably increased, while the share of 'Ukrainian news' decreased. Some of these changes can be explained by typical changes in the news agenda, but more generally we can say that the qualitative and quantitative shifts in the discussion regarding the identified ideological topics indicate users moving to the pro-Russian infosphere.

	BEFORE THE BAN	AFTER THE BAN				
	Share of unique authors, %	Share of unique posts, %	Share of posts among ideological posts written before ban, %	Share of unique authors, %	Share of unique posts, %	Share of posts among ideological posts written after ban, %
Religion	70.42%	88.66%	20.92%	59.29%	91.23%	18.08%
Pro-Russian propaganda	65.63%	89.22%	25.64%	48.29%	82.07%	31.84%
Ukraine news	62.71%	88.12%	13.47%	12.64%	82.85%	6.70%
DNR and LPR	55.80%	91.45%	14.30%	33.67%	82.16%	17.73%
Anti-Ukrainian propaganda	55.30%	91.31%	7.27%	23.47%	83.78%	8.24%

Table 1: Permanent Clusters

A Pro-Ukrainian rhetoric cluster exists before the ban that is absent in the second sample; this can be explained by a lack of pronounced features or by an insufficient text length.

An anti-Semitic rhetoric cluster was identified in the second sample; this may indicate some growth in racist sentiment after the ban, although the topic was not discussed more actively after the ban.This topic became mathematically significant because some previously-included topics disappeared after the ban (e.g. pro-Ukrainian rhetoric). Of the 86 posts in the clustering sample, only a dozen are of interest; they represent a wave of posts about fake news issued by Russian media about the World Jewish Congress.

Our analysis of the most significant permanent clusters ('Pro-Russia propaganda', 'DNR and LPR', 'Anti-Ukrainian propaganda'), shows that the share of unique authors decreased from 58.9% before the ban to 35% after the ban, while the share of unique posts decreased from 90.6% to 82.67% on average.

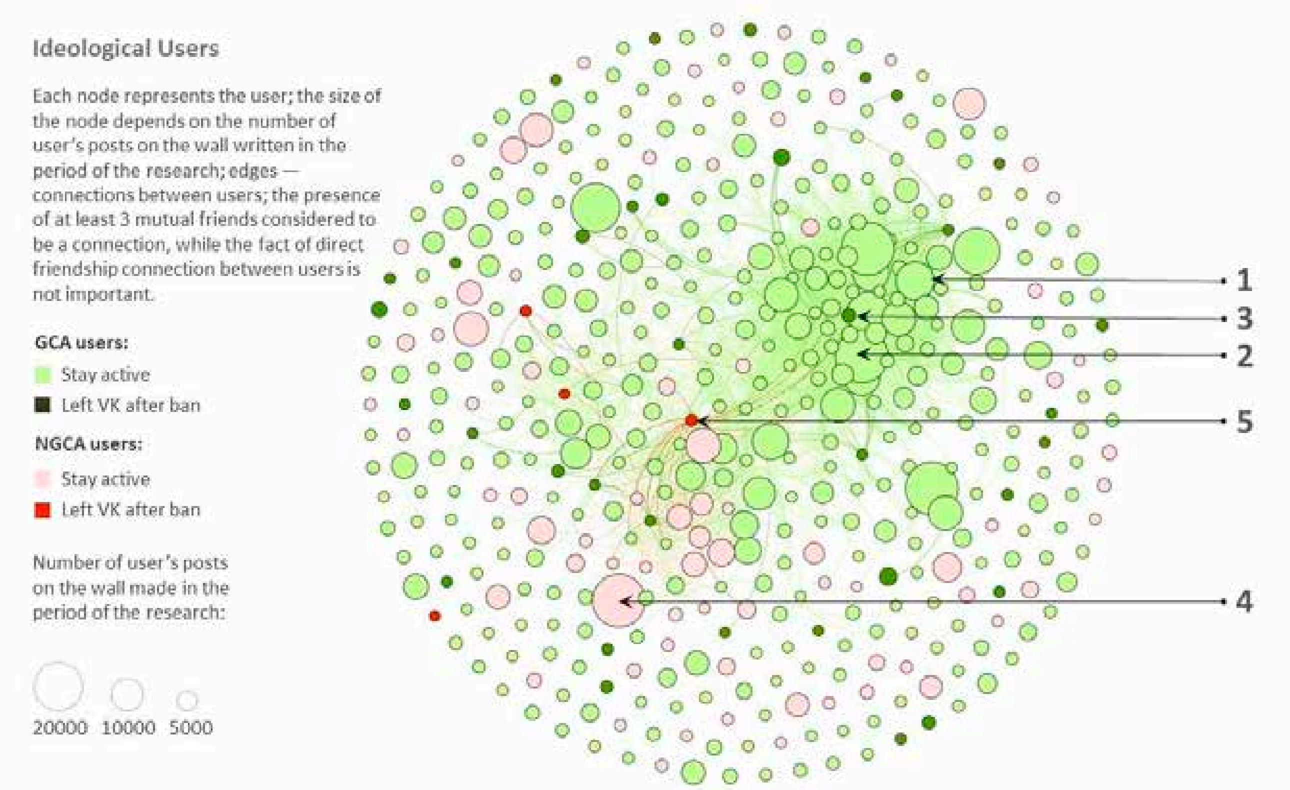

Ideological Users
Each node represents the user; the size of the node depends on the number of user's posts on the wall written in the period of the research; edges — connections between users; the presence of at least 3 mutual friends considered to be a connection, while the fact of direct friendship connection between users is not important.
GCA users:
Stay active
Left VK after ban
NGCA users:
Stay active
Left VK after ban
Number of user's posts on the wall made in the period of the research:
20000
10000
5000
1
3
2
5
4

Figure 15: Connections Among Ideological Users from the GCA and the NGCA

Details about the five representative nodes:

1. This pro-Ukrainian account had 12,985 posts before the ban (or 34.3 posts per day) and 81 posts after the ban. However, all 81 posts were made before 15 May 2018 (or 20.25 posts per day). Despite the fact that the ban was decreed on 16 May, different providers implemented the ban days and weeks later. The profile is located in the GCA, in Eastern Ukraine.
2. This account belongs to one of the most active users in the sample.[82] Posts mostly reveal Russian propaganda. The user remained active throughout all three research periods. The profile is located in the GCA, in Central Ukraine.
3. This was the most-connected account from the GCA region among those that left VK after the ban.[83] It was located in Western Ukraine and the majority of posts were pro-Ukrainian statements.
4. This was the most active account from the NGCA before the ban. The rhetoric used was mostly pro-Russian. Despite an enormous number of posts in the studied period, almost 20,000, the user has only 61 friends, so is unlikely to be an influencer.[84]
5. This account belonged to a commander of pro-Russian military forces in Donbas. According to the graph, the account is well-connected with users in both the GCA and the NGCA. It was one of the small group of NGCA users who ceased activity after the ban. Apparently this was due to the fact that this commander was killed.

The topics 'Anti-Ukrainian propaganda' and 'Pro-Russian propaganda' became more popular together with the 'DNR and LPR' cluster, while 'Religion' lost popularity. The 'News' cluster contained the lowest number of unique posts before the ban. This might be explained by the fact that news is generally reposted without any changes. However, after the ban the 'Pro-Russian propaganda' cluster had the lowest number of unique posts.

Inconstant clusters before the ban: 'Pro-Ukrainian rhetoric', 'Patrol Police', 'Nadiya Savchenko',[85] 'International Monetary Fund (IMF) in Ukraine', and '#KrymNash'. Changing clusters after the ban: 'Russian news', 'Russian presidential election', 'Donald Trump and the USA', and 'Antisemitism'.

We identified user accounts posting about ideological issues before the ban. There were 467 such profiles (or 0.16% of the initial sample) from both regions. Most of these accounts continued to use VK after the ban. Only 35 profiles from the GCA and four profiles from the NGCA left the network (shown as dark green and dark red respectively in Figure 15 below).[86]

As shown by the graph, the characteristics of ideological users differ from those of most other profiles. Let us examine them in greater detail and comparing the ideological user with a typical user from the initial sample of 315,697 profiles.

Unlike the typical user who posted on average once in four days before the ban and once in ten days after the ban, the ideological user was significantly more active generating 4 posts per day before the ban and 1.6 posts after the ban. On the other hand, the activity

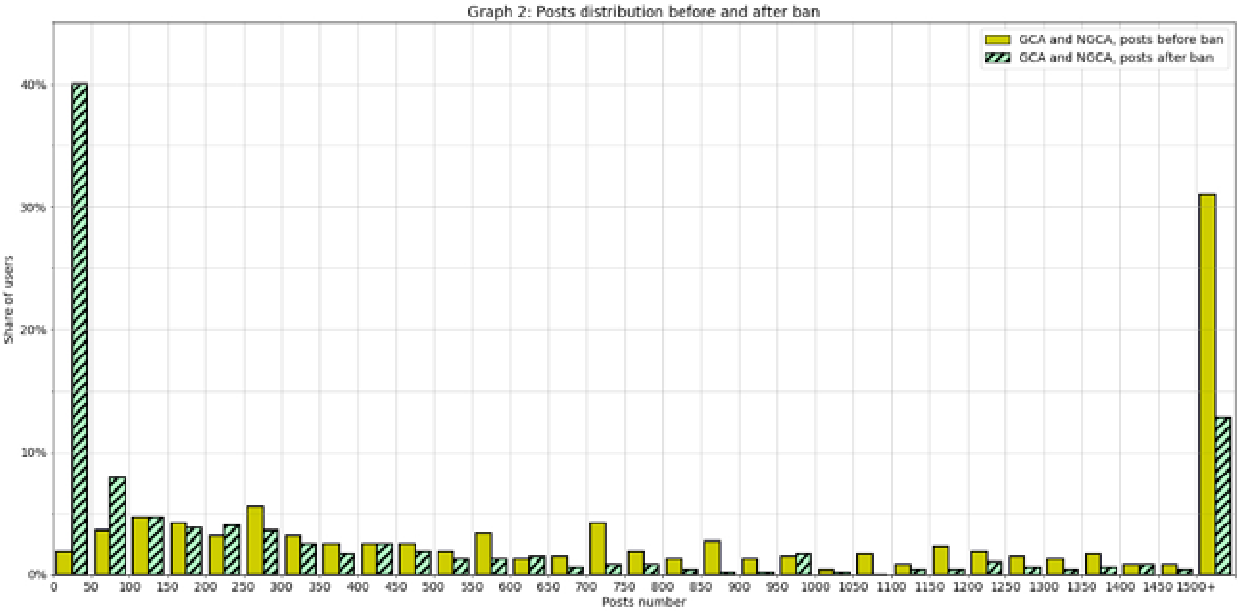


Figure 16: Posts Distribution Written by Ideological Users Before and After Ban

of typical users decreased by 55% after the ban, while the activity of ideological users dropped by 62%. Although the number of those who wrote fewer than 50 posts soared from 2.4% to 40%, the number of those who posted four and more times per day fell by almost 14% after the ban.

Ideological users are significantly more connected than typical users. While the typical user had on average 195 friends before the ban and 217 after the ban, the number of friends for a typical ideological user grew from 197 to 501. The number of users with 1,000 or more friends increased by almost 10 percentage points.

While before the ban the typical VK user consumed information from slightly more groups than the average ideological user (106 vs 90 groups on average), after the ban the situation was reversed. On average ideological users subscribed to 2.25 times more groups than typical users (121 vs 271 groups).

In order to analyse ideologically-tinged traffic generated within groups, we created a list of 444,000 groups, from which ideological users reposted. Among these groups, only 5% had more than 100 reposts.[87] These were short-listed, and a 500 post group was downloaded and labeled using the results of the clusterisation technique described above.[88] From this we calculated the share of ideological traffic. The list of ideological groups was further refined to include only those with at least a 25% share of ideologically tinged posts, resulting in a set of 620 groups selected for further analysis.[89] Some groups were manually excluded as they did not meet the study objectives.

The number of posts[90] published in the ideological groups during the studied period decreased by 16.7% as a result of the ban.[91]

Our analysis shows that the number of reposts from ideological groups generated by Ukrainian users fell by a factor of 2.5 after the ban.[92]

The number of views[93] for posts generated by ideological groups decreased by 19.4%.[94] However, at the end of March and beginning of April 2018 this number returned to pre-ban values. Peaks occur during the presidential elections in Russia (18 March 2018), on the day of a big fire in Kemerovo (26 March 2018); [95] and on the day of the anti-Putin rally (5 May 2018).

In order to discover how pro-Ukrainian communities reacted to the ban, we studied the activity of 39 such groups. We used a list of pro-Ukrainian groups observed by the Russian Security Forces during Ukrainian civil protests in 2013–14.[96]

Among those 39 groups:

- 8 groups are still active.The themes they post about have changed from Euromaidan to war in Eastern Ukraine. 4 of these groups have been blocked from visiting Russian IP addresses.

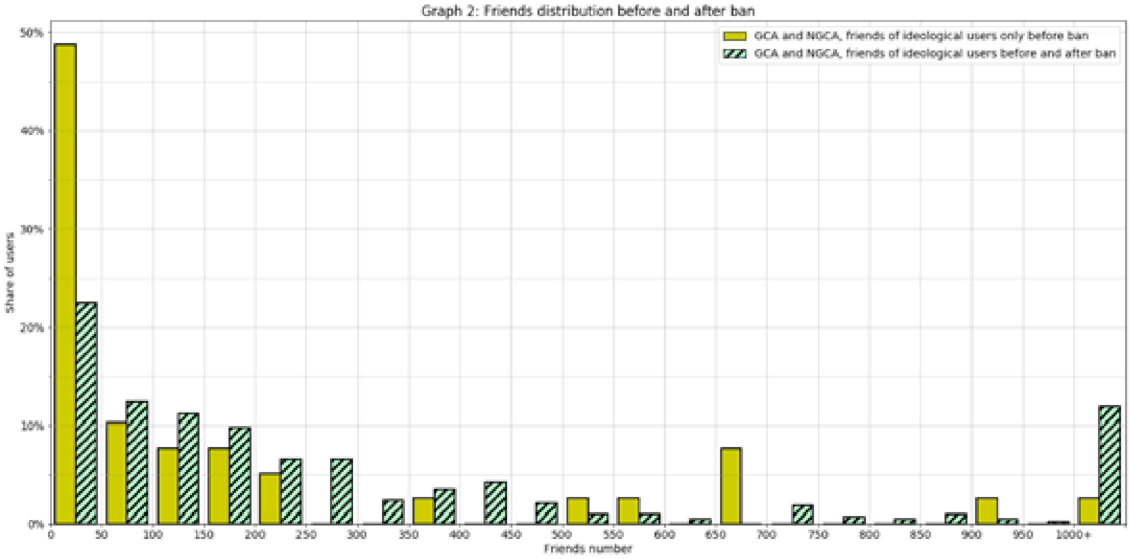


Figure 17: Ideological Users' Friends Distribution Before and After Ban

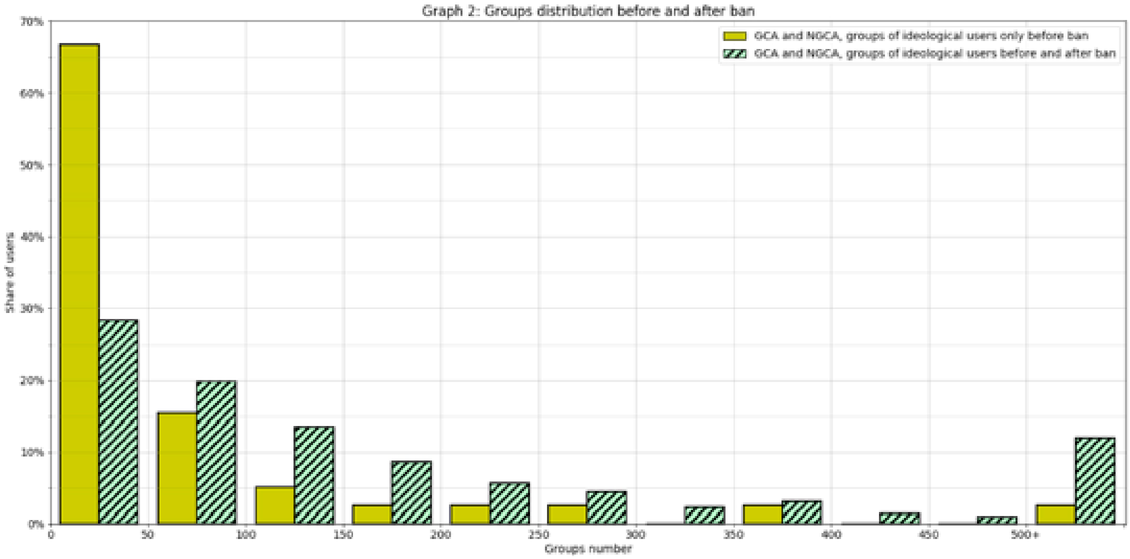


Figure 18: Ideological Users' Groups Distribution Before and After Ban

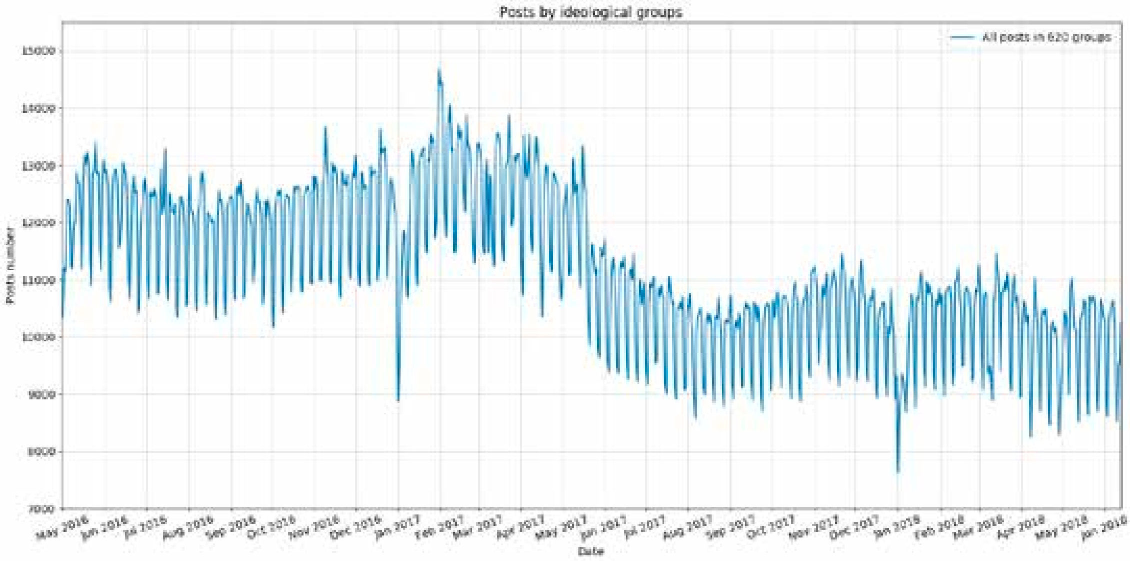


Figure 19: Posting Dynamics Among Ideological Groups During the Research Period

- 20 groups have been deleted/blocked or have radically changed the themes about which they post.
- 4 groups ceased activity even before the ban.
- 7 groups ceased all activity after the ban.

Only 8 out of the original 39 pro-Ukrainian groups are still active, further increasing the prevalance of anti-Ukrainian rhetoric in VK.

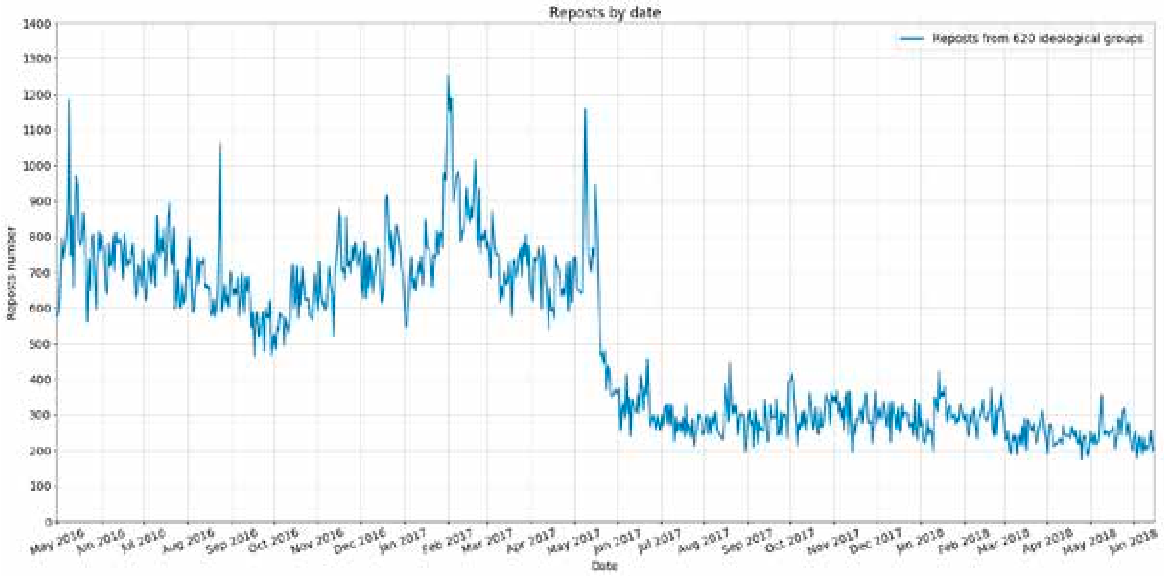


Figure 20: Reposting Dynamics from Ideological Groups by Ukrainian VK Users during the Research Period

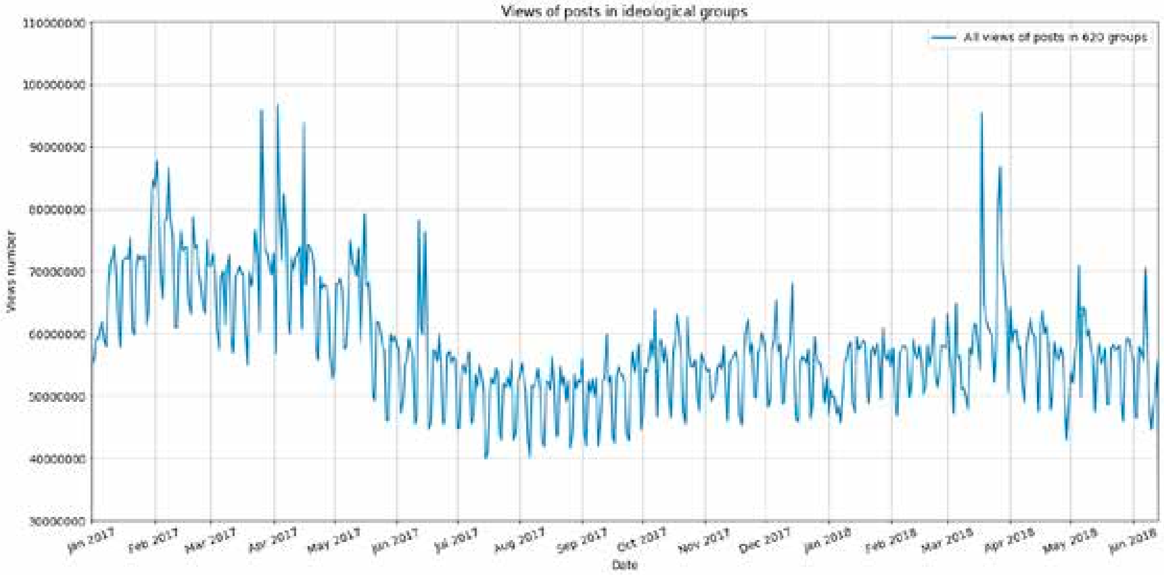


Figure 21: Views Dynamics in Ideological Groups (from the moment of appearance of this functional in VK)

CONCLUSIONS

The ban of the Russian online social network VKontakte provided a unique opportunity for studying the effects of such a ban in the context of hybrid warfare. Was the VK ban effective? As in the case of most complex issues, there is no clear-cut answer, but rather a number of pros and cons.

As stated in the Ukrainian Law 'On Sanctions', the main reasons for the ban were: 'to protect the national security and territorial integrity of Ukraine, and to counter terrorism'. In this regard, the ban can be considered to have been effective.The VK audience diminished and user traffic decreased, so this channel for pro-Russian propaganda directed at broad sections of the Ukrainian population was effectively narrowed.

In fact, the VK audience in Ukraine decreased by a factor of 3.2.[97] However, the dissemination of VK in the GCA territories still under the ban is now only 19% lower than in the NGCA (or territories occupied by Russia). In addition, despite the fact that total VK traffic was reduced by two-thirds, the total number of posts in the GCA is only 55% lower than in the NGCA, and the posting frequency in the NGCA after the ban was only 21% higher than in the GCA.[98]

The profiles that continued to use VK after the ban became more connected. Users residing in territories subjected to the ban have 16% more friends on average than they had before the ban. In addition, users from the GCA tend to consume information from a greater number of groups and now subscribe to 20% more groups than before the ban.

The fact that the VK audience has become younger on average is also troubling. Young people can be assumed to be more easily motivated to action than older people. Users from the GCA who continued to use VK after the ban are, on average, 4.2 years younger than users from the NGCA (27 vs 31 years old).[99] The average age of a typical user from the GCA decreased by 1.7 years compared with the period before the ban.[100]

Despite the fact that the number of ideological users decreased as a result of the ban, those who continued to use VK became significantly more active compared to the average user. Before the ban, a typical user would post, on average, once in four days; after the ban the frequency decreased to once in ten days. In comparison, an ideological user wrote 4 posts per day before the ban and 1.6 posts per day after the ban. The share of ideological posts increased from 36% before the ban to 52% after it; this can be compared with a decrease in reposts from ideological groups from 64% to 48%.

The study demonstrates that pro-Russian narratives are now reverberating throughout the Vkontakte social network. This ‘echo-chamber effect’ potentially increases the risk of radicalisation, and therefore the Ukrainian authorities should consider this situation a security concern. On the other hand, the dramatic decrease in VK use will make it easier to detect vulnerable individuals. In other words, the operation of a VK profile after the ban, which is also self-reporting as located in ‘Ukraine’, indicates the potential for radicalisation.

We conclude that the most pressing current problem with VK is that law-abiding, pro-Ukrainian citizens were among the first to stop using the social network, which meant that those who continued to use VK oppositional voices, and consequently the rhetoric that appears on the site is less diverse, strengthening the echo chamber. Posting frequency in the GCA remains high (1 post per 4 days per average user), so VK remains an important platform for pro-Kremlin voices in the information environment.

While some suggest the ban was effective in supporting the strategic objectives that justified its enforcement, it is still too early to know if the side effects of the ban have the potential to outweigh the positive effects of the ban. From a strategic point of view, it is premature to come to any conclusions about the total outcome of ban. We must remember that information activities do not exist in a vacuum, but represent one of a number of components comprising the hybrid warfare paradigm. As these different components work in concert, it will be necessary to support all lines of effort for Ukraine to consolidate the gains made so far in fighting information activities.

05

DIVISION ABROAD, COHESION AT HOME: How the Russian troll factory works to divide societies overseas but spread pro-regime messages at home

John D. Gallacher & Rolf E. Fredheim

ABSTRACT

In the last 6 months increasing quantities of information have been released regarding exactly how attempts to influence the 2016 U.S. presidential election may have been carried out. In this study we use a combination of quantitative analysis, machine learning and natural language processing to map out the topics of conversation promoted by the Russian troll factory, the Internet Research Agency (IRA) between 2015 and 2018. We show how this activity involves coordinated messaging across multiple social media platforms (Facebook, Twitter, Reddit and Instagram), and how this seeks to promote both sides of controversial debates with inflammatory material. Additionally, we demonstrate how real-world events are utilised to spread division in societies abroad with a common pattern of progressively provocative content. Finally, we show how this same agency uses these platforms for a very different purpose for domestic audiences, and spreads single pro-regime messages without attempts to intensify divisions at home. These results demonstrate how any successful solutions to counter this type of activity will need to tackle the problem from a multi-platform approach, and also must consider how alternative audiences may be targeted in different ways by hostile influence.

INTRODUCTION

For nearly two years, security professionals, politicians, and researchers have been discussing how Russia may have 'hacked' the 2016 US election using a combination of cyber-operations. These operations targeted the physical election infrastructure, the infrastructure of US political parties, and also the minds of the American public with a prolonged information campaign that sought to divide, inflame, and provoke existing social tensions through the use of social media.[101]

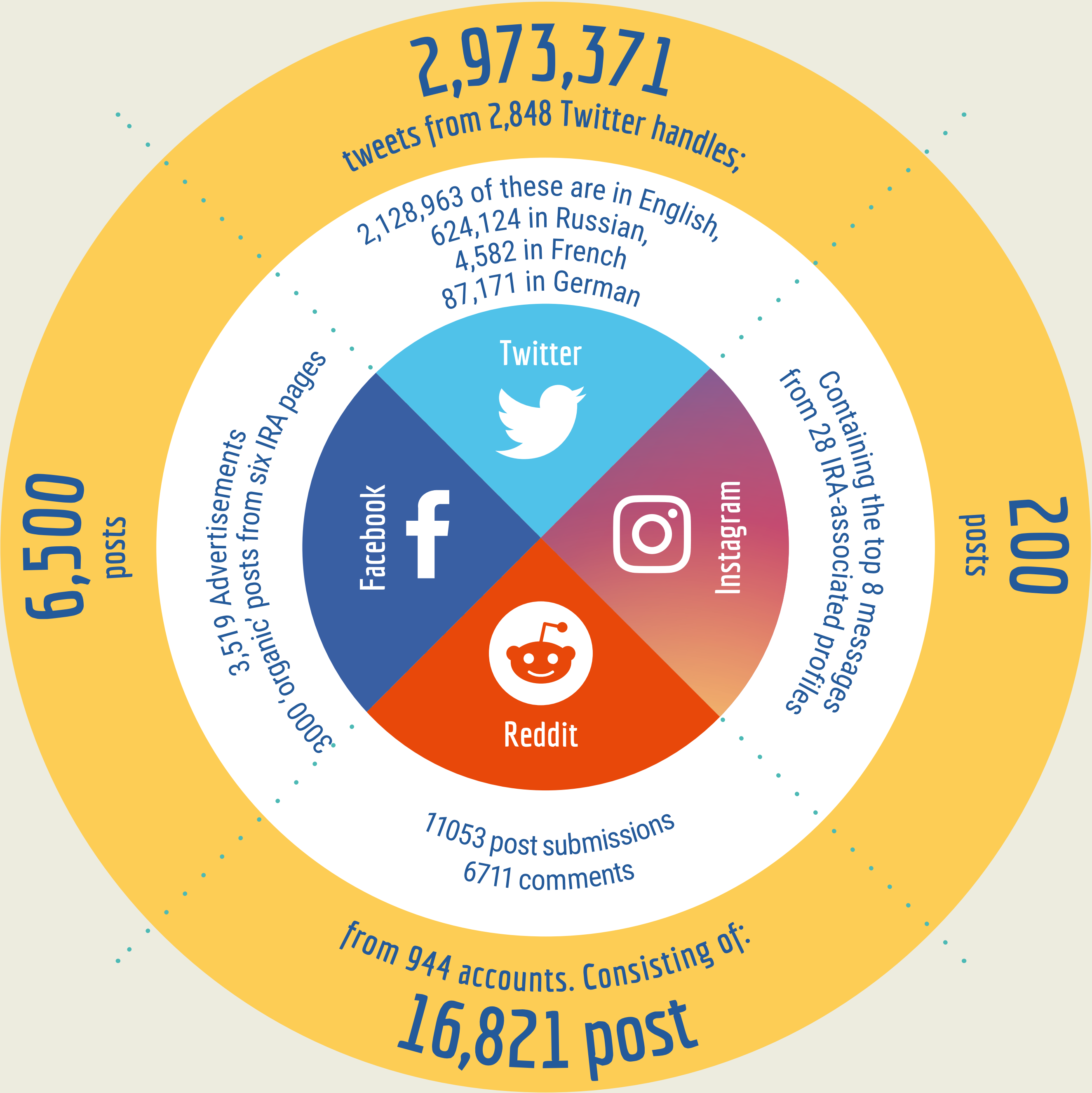
In the last 6 months, detailed analysis of how these information campaigns were conducted has been made possible through the publishing of large open source datasets detailing the posts from accounts attributed the notorious Russian 'troll farm', the Internet Research Agency (IRA). Here we analyse data in both English and Russian across four major social media platforms — Facebook, Twitter, Reddit, and Instagram — to quantify how the tactics differed for domestic Russian audiences and for foreign audiences, and to discover how platform-specific features were gamed for greater effect. The data clearly show these activities were not confined to the 2016 US election, but continued long after, with spikes of activity coinciding with real world events such as the European terror attacks in 2015–2017, right-wing protests in Charlottesville in 2017, and the National Football League (NFL) 'take a knee' protests.[102]

By mapping the change in topics of conversation over time, we show how both sides of divisive debates were stoked using inflammatory content injected into the opposing camp by exploiting specific features of the social media platforms. The tactics and methods used in English demonstrate a high degree of cross-platform coordination designed to spread divisive messaging and polarise both sides of controversial debates in the US and the West. Accounts identified as originating from the IRA sought to pit one half of society against the other by promoting hostile disagreement in the English language space around controversial topics such as racial division, police brutality, immigration, and politics. Conversely, messages in Russian followed a different strategy, pushing pro-regime narratives that mirror those of official state media outlets and rarely act to promote both sides of the debate.

In the second part of the study, we identify a consistent 'troll spin cycle' that occurs in response to real world events. This pattern of activity appears designed to utilise certain types of events to further sow division. We demonstrate how progressively provocative content is injected into the online discussions following large newsworthy events including terror attacks and political protests. We found that this is a common pattern of activity, used to 'weaponise' messages targeted towards the US, the UK, Canada, France, and elsewhere, but that a very different pattern of response is shown toward events in Russia, one that appears to mirror official state narratives.

THE AVAILABLE DATA

Our datasets consist of just fewer than 3 million Twitter messages,[103] 16 thousand Reddit posts,[104] 6 thousand Facebook posts,[105] (approximately 50% organic, and 50% paid advertisements) and 200 Instagram posts.[106] In all cases the accounts that spread these messages were attributed to the IRA by the platforms using proprietary information not available to the public. This information may include



2,973,371

tweets from 2,848 Twitter handles;

2,128,963 of these are in English,
624,124 in Russian,
4,582 in French
87,171 in German

6,500 posts

3,519 Advertisements
3000 'organic' posts from six IRA pages

200 posts

Containing the top 8 messages
from 28 IRA-associated profiles

11053 post submissions
6711 comments

from 944 accounts. Consisting of:

16,821 post

metadata such as IP-address, location, language setting, previous account history, etc. These data represent only a small sample of the total activity conducted by the IRA, but nevertheless provide unique insight into how these accounts were operated and the aims they sought to achieve. While the total reach of these posts is hard to gauge, estimates suggest that at least 126 million people may have seen Facebook posts linked to the IRA,[107] 145 million users may have seen the posts on Instagram,[108] and 1.4 million users on Twitter were exposed to the content.

With the exception of Reddit, these data were removed from the platforms at the point of attribution and are not available for public analysis. The datasets we use here have come from independent researchers who collected the data prior to its removal and then made it public. We thank everyone involved in making this data available. This study would not have been possible without it.[109]

1 TOPIC LANDSCAPE

1.1 Cross-platform coordination spreading division abroad

Our analysis shows a high degree of cross-platform coordination, ensuring the same controversial topics spread across Facebook, Twitter, Reddit, and Instagram in English. A combination of machine learning and natural language programming allowed us to classify each social media post according to the related topic of conversation, and then plot how the activity patterns for these topics changed over time. Five broad topics of conversation emerged — social issues, Syria and international terrorism, race issues, overtly political content, and 'other'. The activity patterns for the four platforms are shown by topic in Figure 1.

In order to visualize how much of each conversation addressed opposing sides, we took a sample of each topic and manually coded it according to whether the message addressed one side of the debate, the other side of the debate, or took a neutral stance. We also measured the average 'toxicity' of each topic, using the Google Perspective API. This classification tool uses machine learning models to score the perceived impact a comment might have on a conversation. Comments that are defined as being rude, disrespectful, or unreasonable receive a higher 'toxicity' score (0 = low toxicity, 1 = high toxicity). *[See the Annexe for further details on these classifications.]*

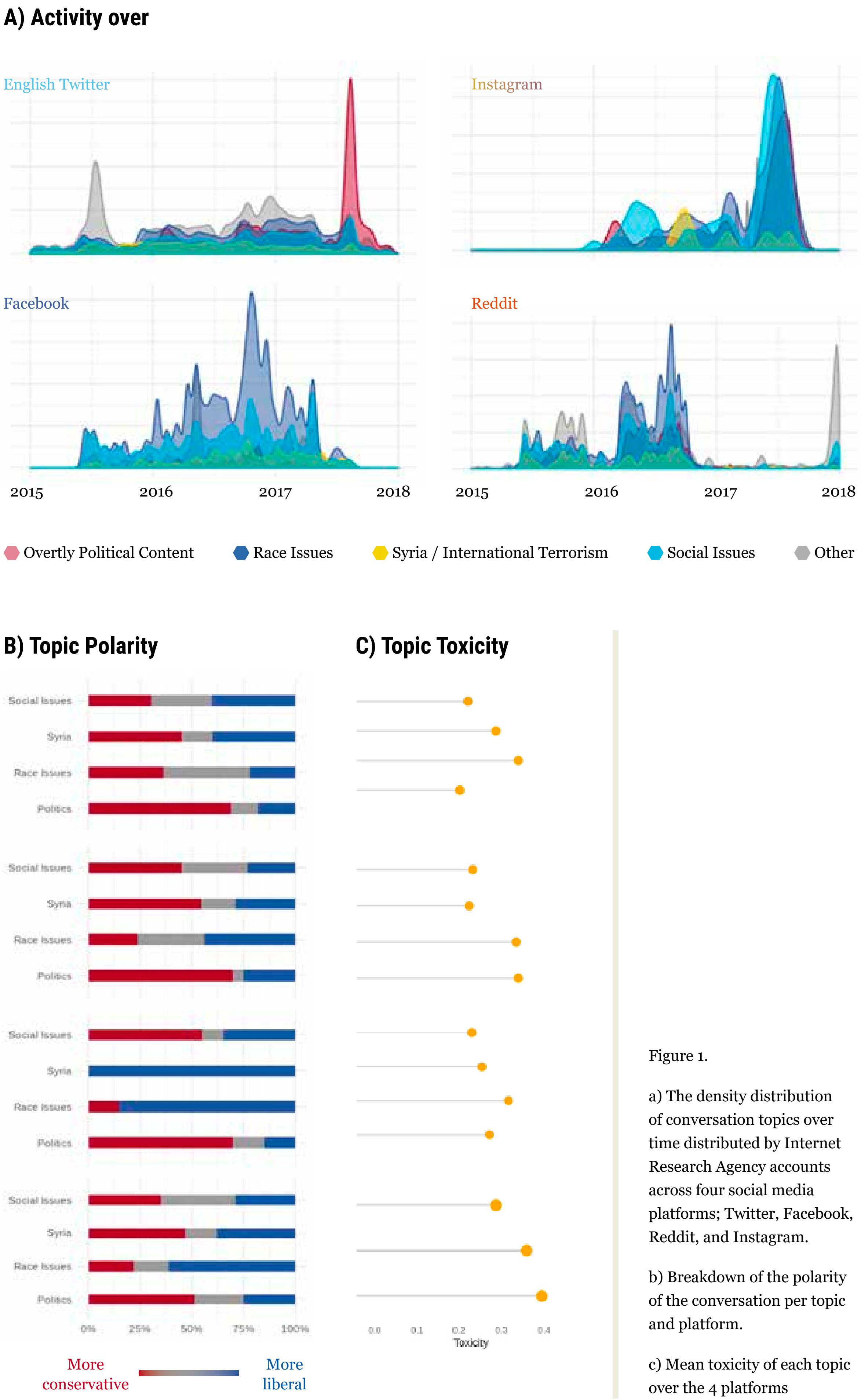


Figure 1.

a) The density distribution of conversation topics over time distributed by Internet Research Agency accounts across four social media platforms; Twitter, Facebook, Reddit, and Instagram.

b) Breakdown of the polarity of the conversation per topic and platform.

c) Mean toxicity of each topic over the 4 platforms

1.1.1 Reddit

Issues relating to US politics and racial tensions dominated conversations on Reddit. Messages in support of the Black Lives Matter campaign and decrying police brutality dominated, as the accounts attempted to widen existing divisions within communities, and to increase inter-group tensions and conflict. The Reddit dataset also contained conversations about civil unrest, the Syrian conflict, international terrorism, and financial news, together with 'other' non-political content which appears designed to create more realistic account histories and gather higher 'karma scores', appearing as valued members of the site. Known IRA activity peaked in mid-2016 and has tapered off since, with recent spikes in activity addressing non-political topics, including the promotion of bitcoin and other cryptocurrencies.

1.1.2 Twitter

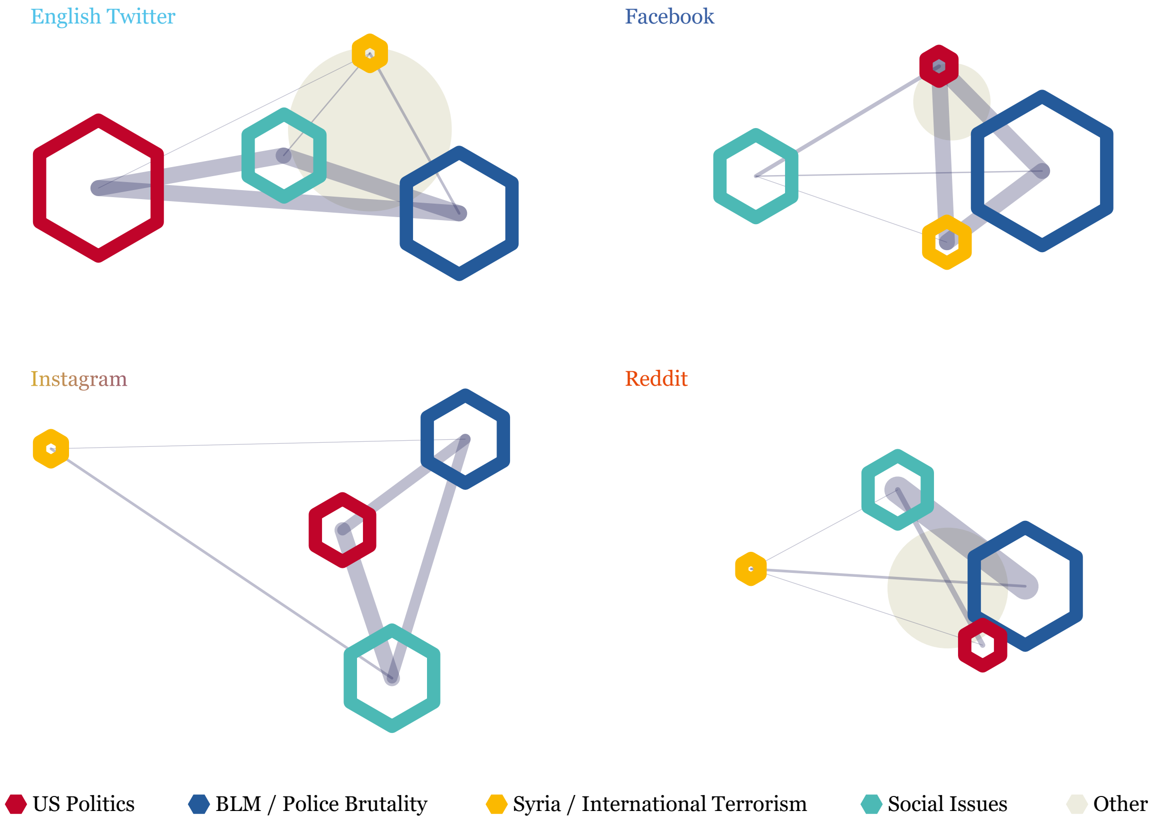
English-language Twitter posts were concerned with the same topics as those on Reddit, but the proportion of messages about US politics was greater than for racial issues. These topics were also less closely linked. Political commentary included messages in support of President Trump and the Make America Great Again campaign, but also in support of other candidates in the US primaries including Hilary Clinton and Bernie Sanders. On the other hand, material about the Black Lives Matter movement picked up on broader issues likely to appeal to liberal and Democratic-leaning audiences. More recent activity has tended towards political discussions, with a spike of right-wing troll activity in late 2017.

1.1.3 Instagram

The conversation on Instagram was dominated by discussions of social issues including veteran's affairs, gun control, LGBT rights, confederate iconography, and religious issues (covering both Christianity and Islam). Most striking within the Instagram dataset was the upward trend in activity over time, suggesting that this platform is being targeted to a greater extent as the popularity of the platform increases. Although the quantity of data available for Instagram is much lower than for other platforms, the breadth of topics targeted and the complexity of the posts suggest that significant effort went into creating these posts, and the data we do have is therefore likely part of a much larger campaign.

1.1.4 Facebook

Discussions of racial tension dominated the activity on Facebook, speaking to both the Black Lives Matter movement, and against the activity of right-wing groups. They also addressed topics of police brutality, veterans' affairs, and the second amendment to the United States Constitution. Throughout this dataset there is a constant lower-volume backdrop of activity regarding the threat of international terrorism and discussions about immigration. There is a notable lack of directly political content in this dataset,



however it is important to note that these data represent only a small sample of the total IRA activity occurring on Facebook. (Data are only available for six of 470 identified IRA-created Facebook accounts,[110] and 88,000 organic posts from these accounts are yet to be made public). We are not aware of the criteria used to select these posts for release, and it is possible that more directly political content is yet to be made available.

1.2 Polarisation and Toxicity

In addition to the controversial nature of the topics themselves, the IRA troll farm supported and antagonized both sides of these debates. We found evidence in all topics that both opposing sides were being promoted. See Figure 1(b). Notably, while we classified many posts as 'neutral' due to non-partisan themes in the messaging itself, many of these posts spread news of Western terror attacks, violence, criminality, or unrest. These posts appear to have been selected to give a broad sense of instability, even if they are not directly related to a specific political party, statement, or ideology.

The toxicity of posts was coded for all platforms and topics. See Figure 1(c). Posts relating to race issues were consistently scored as the most toxic, implying that such content was the most likely to inflame, agitate, and generate hostility. The observation held true across all platforms, again suggesting a certain degree of coordination. Google's Perspective API rated all topics as fairly toxic; across all platforms, the average toxicity score was 0.31. By comparison, a sample of generic Tweets from non-IRA Twitter activity (taken from the BBC News Feed) recorded an average toxicity rating of just 0.12.

1.3 Relationship between topics

An overview of the relationship between topics is shown in Figure 2. The topics — represented by circular nodes — are scaled to reflect the number of posts about each topic, while the connections between topics map the frequency at which topics occurred together. It is clear that while the topics are similar across all four platforms, each platform displays certain unique characteristics. On Reddit there was a higher degree of connection between political conversations and conversations about racial issues, suggesting that these topics were often discussed together, with race issues used to make political statements. Conversely on Twitter, the political node is more closely related to broader social issues. Topics on Instagram and Facebook tended to be more distinct, which is reflected in a greater distance between the nodes and weaker connections. This suggests that messages on these platforms more often addressed a single topic.

1.4 Platform Features

Platform-specific features further demonstrate how certain conversations are being co-opted to increase aggressive inter-group contact. We investigated hashtags within the Twitter dataset and found that hashtags supporting opposing sides of a discussion would frequently be used in conjunction within a single Tweet.

For example, the #blacklivesmatter hashtag was often used in conjunction with the #whitegenocide hashtag, a movement that argues strongly against racial diversity and makes conspiratorial claims about white oppression in the United States. Equally, #blacklivesmatter was spread in conjunction with the #alllivesmatter hashtag, bringing together two opposing sides of a racial divide. #imwithher, a pro-Hilary Clinton movement, often co-occurred with #crookedhilary, while opposing sides of the debate were also joined together for #policebrutality, #bluelivesmatter, and #policelivesmatter, and for #MAGA a movement supporting President Trump along with hashtags opposing the presidency, such as #resist and #notmypresident.

Similar trends were also demonstrated on other platforms — cross-posting of counter-attitudinal messaging took place between specific communities (sub-Reddits) on

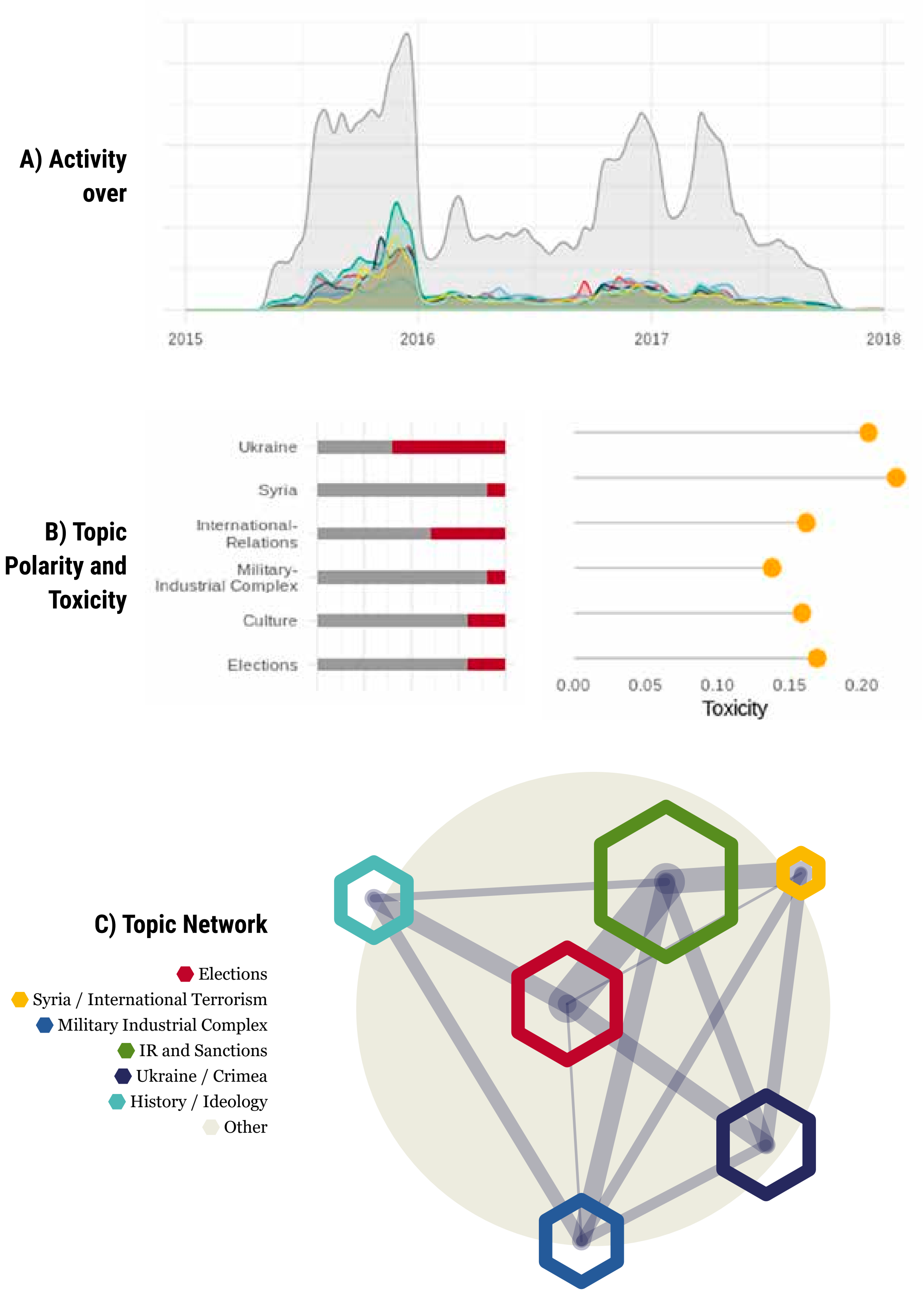
Reddit, and out-group members were often targeted with counter-attitudinal messaging on Facebook. For example, adverts targeted pro-immigration groups with messages promoting stricter border control and conservative groups were shown messages addressing liberal topics.

1.5 Cohesion at Home

Compared with the topics present in English-language messages, messages aimed towards Russian-language audiences differed sharply. See Figure 2. While the English language content appears tailored to provoke division, Twitter activity promoted to domestic, Russian-speaking audiences is much less divisive. Accounts posted in support of the ruling party, praising the regime's position on Syria and Ukraine, and exaggerating divisions and threats in the West. These accounts rarely discussed highly controversial topics in the Russian language space, or sought to inflame both sides of the debate. The conversation was dominated by Russian international relations, the conflict in Ukraine, and political posts relating to internal Russian politics. These topics do not feature prominently in English-language conversations. The only point of convergence across all both languages is the conflict in Syria and the threat of international terrorism. Within Russian Twitter the largest topic was content that we classified as 'other'. This content related to the entire spectrum of Russian news — from sport and weather, to entertainment and business. The broad character of these subjects reflects that the Kremlin-supported trolls and bots seem to be indiscriminately promoting news content from state media outlets rather than tailoring specific content for Twitter audiences.

How these topics developed over time and the level of polarization is shown in figure 3a-b, with the relationship between topics shown in figure 3c. The IRA's activity in Russian shows sensitivity to breaking news stories and the ability to rapidly switch focus. Over time messages on Russian-language Twitter have trended away from international questions in 2015 to become more internally focused. In particular, messages drawing on historical and ideological themes have emerged. Compared to English language activity, Russian language was much less polarised, and very rarely provoked both sides of controversial topics. This is shown in figure 2(b), with the majority of the content taking a neutral stance (shown in grey), while a smaller percentage took an emotive pro-government position (shown in red). There was no activity within our sample that took an opposing anti-government position.

Overall Russian Twitter Toxicity scored on average 0.18. This is much lower than the English language content, 0.31, and therefore less likely to inflame and increase inter-group tensions. In Russian the most toxic topics were those relating to the conflicts in Syria and Ukraine, implying that these topics were the ones that invoked more emotive language.



A) Activity over

B) Topic Polarity and Toxicity

C) Topic Network

1.6 English language customization

Patterns in platform usage also reveal that a higher level of customization and targeting has been invested in English-language activity, compared to activity in the Russian-language space.

Compared to Russian-language Twitter posts, English-language posts are less likely to contain external links, more likely to be directed at other users, and also more likely to contain hashtags. This demonstrates how messages in English actively sought to build an audience within Twitter and engage directly with other users. Russian-language messages, by comparison, mainly distributed external news content and promoted messages taken from state-run media. In Russian, Twitter appears to have been used to game platform metrics, such as 'most shared' news story and video rankings, rather than to engage with users directly.

Metric	English	Russian
Links to external sites	64%	90%
Mentions of other users	15%	6%
Hashtag use	46%	15%
Highly specialized accounts	34%	15%

Users active in English-language spaces also exhibited greater focus in their topics of conversation. We labelled accounts commenting on a single topic more than 50% of the time as 'highly specialized'. Within English-language Twitter we found 34% of accounts were highly specialized, compared to only 15% of Russian-language accounts. This suggests that English-language accounts seem to be tailored to specific groups, and the accounts are used for individual and specific purposes, while in Russian the same accounts speak to a much broader audience.

2 THE "TROLL SPIN CYCLE"

2.1 Response to real-world events

This difference between the IRA's activities spreading division abroad and cohesion at home can be further demonstrated by comparing how the organisation's Twitter activity in English and Russian responded to real-world events, such as terror attacks, political protests, and international military conflicts. We found a common pattern of activity in the messages targeted towards the US, the UK, Canada, France, and elsewhere in the West, but a very different pattern appeared in response to events in Russia.

As the available Twitter dataset contained the greatest amount of data and covered the longest time period, we used it for the following analysis. We sampled activity on Twitter following seven real world events that occurred over the last three years, five of which were took place in Western countries, and two of which occurred within the Russian sphere of influence:

- 'Unite the Right' Rally in Charlottesville in August 2017
- Paris terror attacks in November 2015 (figure 5)
- March 2016 Brussels bombings
- London terror attacks in March and June 2017
- Quebec City mosque shooting in January 2017
- The 2015 Russian Sukhoi Su-24 shootdown
- The March 2017 Russian Protests

The sampled Twitter data demonstrated a consistent pattern of activity in English, and in local languages, for tweets concerning the Western events. Immediately following one of these events, the IRA's first would first distribute news content in the form of URLs and hyperlinks to draw attention to what had happened. Then their messages turned to expressions of sympathy and concern, possibly to appear 'local' and access the wider conversation. The next tactical step in the pattern was a switch to more hostile rhetoric, including the use of out-group messaging (speaking in terms of 'us' vs. 'them'), and connecting the event to broader social issues such as immigration and wider political discussion. Finally, the rhetoric incorporated conspiracy theories, spreading progressively wilder theories to retain attention and further spread division among those who might be vulnerable to such messaging. See Figure 4.

Importantly, this cycle was absent from the Russian Twitter data. Following both Western security events and Russia-specific events the IRA's messaging stayed 'on point' for the entire period. Emotive language was used to highlight elements of betrayal, but the type of messaging did not vary over time. While there were mentions in Russian Twitter of the terror attacks that occurred in the West, they dropped out of the dataset approximately one week after each event — there was no lingering or weaponisation of these event for political gain. For example, tweets relating to the 2015 Paris attacks ended within five days of the event in Russian Twitter, but lingered on much longer in French and English.

The same reflection of state-run media messaging was also present in Russian Twitter activity related to the downing of a Russian Sukhoi Su-24M aircraft near the Syria–Turkey border on 24 November 2015. The Russia Defence Ministry denied

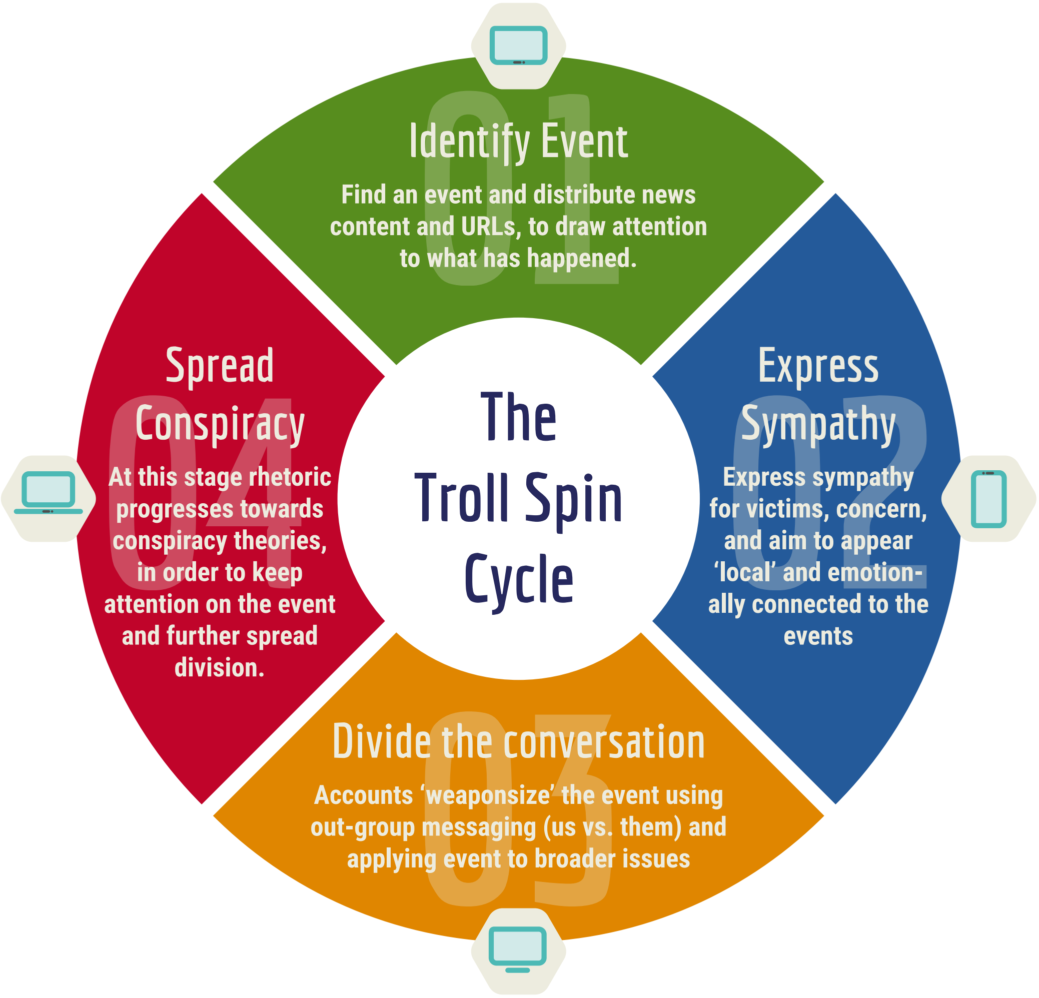
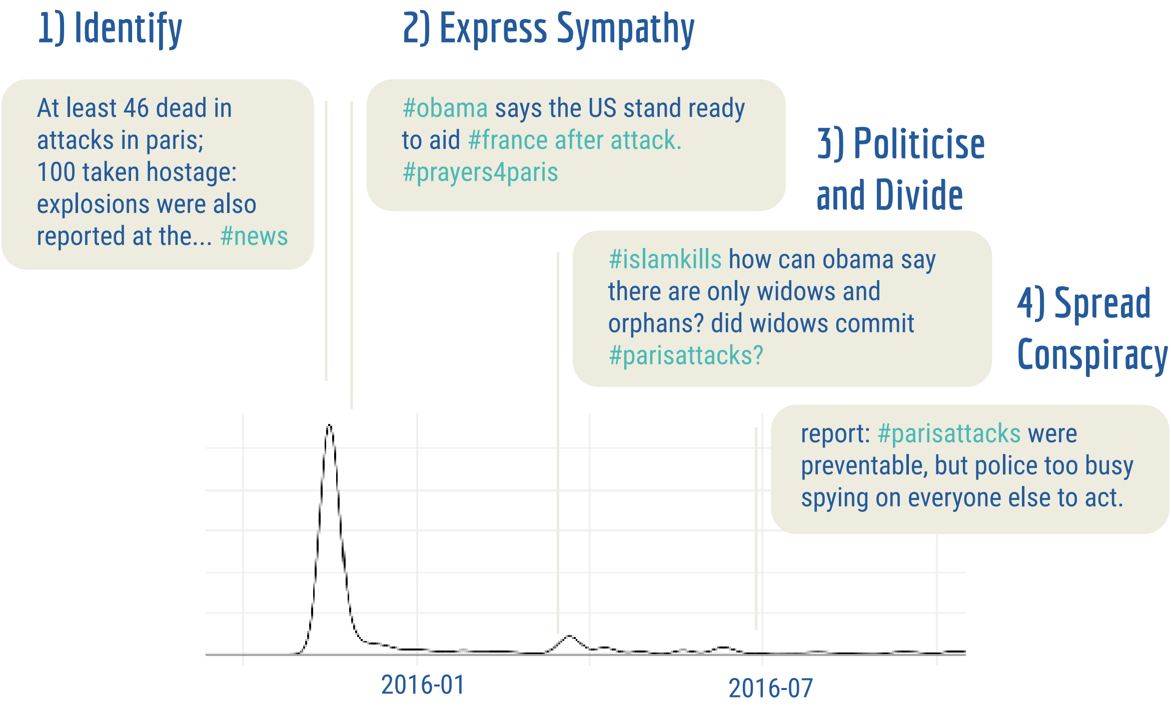

01
Identify Event
Find an event and distribute news content and URLs, to draw attention to what has happened.
02
Express Sympathy
Express sympathy for victims, concern, and aim to appear 'local' and emotionally connected to the events
03
Divide the conversation
Accounts 'weaponsize' the event using out-group messaging (us vs. them) and applying event to broader issues
04
Spread Conspiracy
At this stage rhetoric progresses towards conspiracy theories, in order to keep attention on the event and further spread division.
The Troll Spin Cycle

Figure 4. How English language Russian Internet Research Agency Troll accounts respond to real world events with a consistent pattern of activity.

the aircraft ever left Syrian airspace, whilst according to Turkey, the aircraft was fired upon while 2km inside Turkish airspace. Twitter activity mirrored the Russia's official position, stating that the plane was inside Syrian airspace and that Turkey was the aggressor in this incident; these tweets also promoted hashtags against travel to Turkey and hashtags claiming Turkey had betrayed Russia. However, the Twitter activity did not play to sides of the debate or provoke discussion, but rather expressed a single unified viewpoint. A further example comes from the IRA's Twitter responses to large demonstrations. Again, the activity mirrored Russian state media by ignoring the events. While the Charlottesville 'Unite the Right' rally that occurred in August 2017 generated a great deal of activity in the UK dataset, the Russian protests from March 2017 (which drew larger crowds) are not mentioned in the Russian dataset. On the initial day of the protest only two messages in the Russian dataset referenced the events, in both cases displaying very neutral language. This lack of coverage again mirrors the official response; Russian state television completely ignored the protests, whilst Pro-Kremlin newspapers were equally silent.

2.2 The Spin Cycle in Action – #Parisattacks

Figure 5 shows the troll spin cycle in action following the Paris attacks in November 2015. The initial response from the accounts affiliated with the Russian Troll Farm was to draw attention to the event and spread raw information about the attack. The messages then quickly turned to sympathy, using the #prayers4paris hashtag to join the global conversation. Then the events were politicized, using mentions of the attacks to influence discussions on US politics and religious tensions. Finally, the IRA's Twitter activity continued to linger long after the event, spreading conspiratorial messages and discussing issues such as the attackers' identities and police culpability.



At least 46 dead in attacks in paris; 100 taken hostage: explosions were also reported at the... #news

#obama says the US stand ready to aid #france after attack. #prayers4paris

#islamkills how can obama say there are only widows and orphans? did widows commit #parisattacks?

report: #parisattacks were preventable, but police too busy spying on everyone else to act.

Figure 5. Example of English language Twitter activity following the November 2015 Paris terror attacks showing the spin cycle in action. Troll Farm Tweets transitioned from news content and sympathy, to polarising content and conspiracy theories.

3 HOW EFFECTIVE IS THIS ACTIVITY?

It is important to determine whether or not this activity is actually effective in increasing polarization in the West. We can't answer this question using Twitter data alone as, despite the large number of people who saw and interacted with the content, we cannot directly measure if the tweets elicited a cognitive response in those who read them. However, the Social Identity model of Deindividuation Effects (SIDE)[111] used by social psychologists suggests that divisive activity on social media is able to cause changes in social perceptions and influence offline behaviour.

The SIDE model posits that online environments typically support a high degree of individual anonymity — the personal identities of users taking part in online conversations are often masked. At the same time there is also a high degree of salience in-group identity and group membership of those involved — the social groups that a user belongs to are often more obvious than individual traits. Assumed group identity for the IRA accounts is made even more obvious through the use of distinct imagery such as flags, common iconography, and a high degree of commitment to a single topic. Even usernames reflect certain social attitudes — for example, group identity is written into the username for the infamous @10_GOP account, which posted in support of the US Republican Party.

This combination of high individual anonymity, but high saliency of group membership makes the difference between engaging with members of one's in-group and out-group very striking — it becomes quite obvious when a user is addressing a in-group member with whom they share views, and when the user is addressing an out-group individual. Such 'depersonalization' can lead to a heightened awareness of the crowd, encouraging users to share the crowd's identity and behave in ways that benefit the — for example, over-emphasizing group norms and stereotypes, rather than paying attention to individual traits.[112] Such an environment has differential effects on users who see themselves as members of different social groups and creates hyper-obvious 'us' vs 'them' contact situations. In the dataset we studied, such behaviour appeared across all topics and platforms; political conservatives vs liberals, African Americans vs white supremacists, veterans vs pacifists, and so on.

Once these hyper-salient groups are created, inter-group contact is then artificially encouraged through the use of platform features — cross-posting of hashtags on Twitter and Instagram, invasions of sub-communities on Reddit, and the deliberate targeting of out-group members with divisive advertisements on Facebook. All of these tactics are in play in the datasets. and they successfully led either to direct engagement with out-group members from the artificial accounts, or encouraged genuine social media users to engage with each other, but in a way that is unlikely to be productive – the situation is designed to be confrontational from the outset. Such cases of negative online inter-group contact have been shown to lead to political polarization as each group grows more entrenched in their pre-existing views,[113] and in some cases online polarisation can lead to violence in the real world.[114]

DISCUSSION

We found a high level of cross-platform co-ordination among Russian Internet Research Agency activities in English. This cross-platform co-ordination suggests that any potential response may need to be equally cross-platform in their efforts, and the cooperation of social media companies will be necessary to increase the challenge. It has become essential to follow hostile behaviour across different platforms to design a response with the potential to be effective, whether it involves suspending a fictitious user across multiple platforms to prevent cross-platform pollination of the same inauthentic and disruptive content, or preventing the use of one platform to game the metrics on another. Equally, those conducting less direct counter-measures, such as fact-checking and organic counter-narratives, should also consider focusing on cross-platform responses so as to be able to match the speed of cross-platform disinformation propagation.

Discrepancies between content produced for English- and Russian-language online spaces reflect how greater effort appears to have gone into manipulating the English-language conversation. This high level of individual tailoring – how the platforms were used and which topics were selected – indicates that the stakes were higher in English, the operation more ambitious, and the interventions less crude. This may also indicate that the Russian-language space was easier to manipulate, or that other platforms received greater attention. This latter suggestion reflects previous evidence that social media platforms appear to have done less to police the Russian-language space in comparison to the English-language space,[115] and therefore manipulation in Russian can be more blatant and simplistic.

Our research demonstrates that IRA activity was notably divisive in English, but showed a very different pattern in Russian. It did not seek to promote Russian interests or Russia's standing as a state directly, but instead sought to sow discord in overseas populations and drive groups apart. These efforts may well have been successful.

The manipulation of social media was not restricted to the 2016 US election, and the use of tactics has not gone away. It is vital that efforts to identify malicious activity and construct counter measures should not focus exclusively on election periods, but also consider the damage that can be done by long-term social polarisation in between election cycles. While we cannot currently quantify the direct effect that such activities have had on real users, the SIDE model demonstrates a plausible psychological mechanism through which such online activity can succeed in exacerbating divisions among social groups. This potential dangers of successful manipulation of deindividuation effects in a hyper-polarised online environment highlight the importance of platform responsibility and continued research on this topic. Employing methods derived from social science research in the future may give rise to solutions with the potential to increase the health of social media platforms and make them more resilient to manipulation.

ANNEX: “TOXICITY ANALYSIS”

Topic Model Structure

Figure A1(a&b) shows the topic model structure for English- and Russian-language online social media platforms. These figures break down the identified topic categories into sub-topics. Node size is determined by the relative frequency of each topic in the original dataset across all platforms in English, and for Twitter in Russian.

Topic Polarization Coding

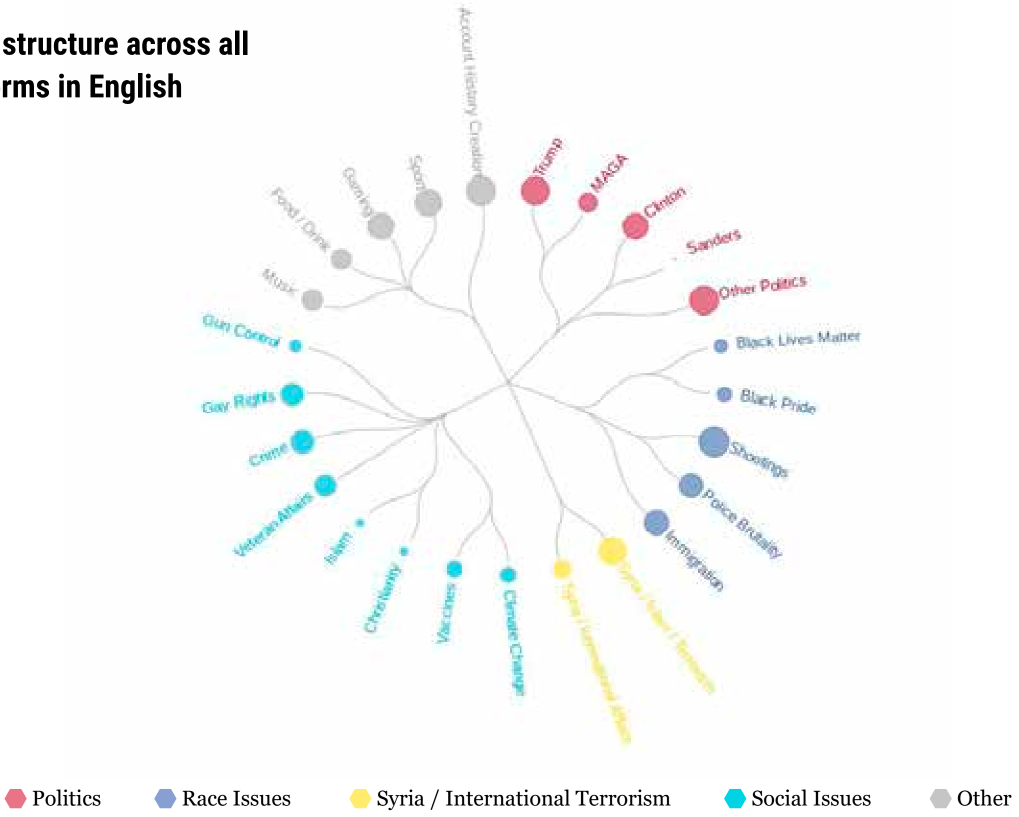
In order to identify how content within each topic was used to inflame both sides of opposing debates we took a sample of messages from each topic and manually coded which broad audience the message appeared designed to address.

For the politics topic this divide is broadly Liberal vs Conservative, while for the Syrian and International Terrorism topic these sides were in favour of and against western intervention. For race issues these sides take pro-establishment and anti-establishment standpoints while broader social issues are classified into progressive vs. reactionary camps.

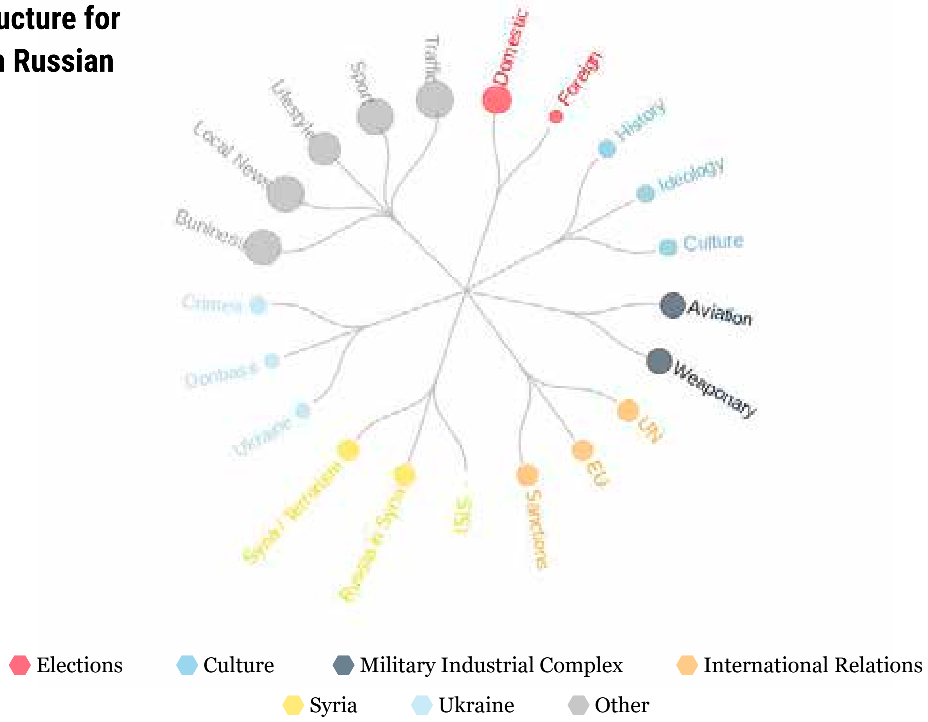
Toxicity Analysis

We used the Google Perspective API[116] to measure the “level of toxicity“ within the

Topic structure across all platforms in English



Topic structure for Twitter in Russian



online conversations. This is a classification tool designed by Google's 'Project Jigsaw' and 'Counter Abuse Technology' teams with the aim to promote better discussions online[117]. The classification tool uses machine learning models to score the perceived impact that a comment might have on a conversation, with comments that are defined as being ruder, more disrespectful, or more unreasonable being more likely to receive a higher 'toxicity' score. The tool was developed through the manual coding of millions of comments from different publishers on a scale from 'very toxic' to 'very healthy'. These resulting judgments provided the large training set of data that the machine learning model was built from.

The model gives a toxicity score for each comment on a scale ranging from 0 to 1. Please note that in some situations, including bad spelling or sentence manipulation, a motivated attacker has easily fooled the Google Perspective API.[118] However, comment abuse detection with deep learning has proved successful in many situations[119] and so when taken at the aggregate conversation level such systems can give a valuable insight into the general nature of the conversation. In the current study each comment was run through the Perspective API using a python script, and from this the average toxicity rating for each event page calculated.

This analysis is currently only available in English, and so in order to get toxicity scores for Russian language we used the Google Translate API to translate the Russian language Tweets into English. Automated translation is often imperfect, and so in order to test that this process did not skew toxicity scores we performed a sense check with a sample of English language tweets. For each of these tweets we obtained the toxicity scores in English, and then translated the tweets into Russian and back into English, re calculated the toxicity scores and compared the two scores, testing for differences. We found that translation did not substantially alter the toxicity results.

06

COUNTERING SUBVERSION ONLINE: what role for public policy?

Edward H. Christie[120]

ABSTRACT

This final chapter aims to offer a synthesis of the main vulnerabilities that liberal democracies contend with, as they encounter contemporary forms of political subversion, and to propose a set of policy principles to guide ongoing reflections on how to best respond to that challenge. Four areas of vulnerability are identified, namely individualised political messaging; group dynamics and political polarisation; platform algorithms and self-radicalisation; and falsehood dissemination dynamics. In discussing each of these areas, insights are drawn from both very recent and more established academic research, at the crossroads of psychology, social psychology, communication studies, and political science. This leads to framing elements for the formulation of proposed policy principles, followed by examples of recent measures in selected countries.

1. INTRODUCTION

The information space that is used by voters, politicians, and interest groups in individual Allied nations is contested and challenged by new risks and threats, both from within and from without. The promise of greater democratic participation and pluralism, provided for by ubiquitous internet platforms such as Facebook, Twitter, and YouTube, has been tempered by concerns about the misuse of personal data and by new forms of political polarisation. This has occurred in tandem with an erosion of the balancing effect of trusted sources of information, and a steep rise in the production and dissemination of false news ("fake news"). The focus of this chapter is on the threats posed to the normally intended functioning of democratic political systems by hostile actors who seek to subvert them for political and/or strategic purposes.

Systemic vulnerabilities in Western political information spaces have been avidly exploited by the Russian state, through the deployment of intentionally divisive and polarising false-flag content, including disinformation, which is defined as "deliberately distorted or manipulated information that is leaked into the communication system of the opponent, with the expectation that it will be accepted as genuine information, and influence either the decision-making process or public opinion".[121]

While hostile non-state actors have also exploited such vulnerabilities, in order to spread extremist narratives and support their recruitment drives, a recent study by two French governmental research institutes reports that an estimated 80% of hostile foreign political influencing efforts in the European Union could be attributed to the Russian Federation[122], and just 20% to other states and to non-state actors combined. For the specific case of the 2017 French presidential election, the study notes that *all* of the foreign influencing efforts that were detected were from Russia.

Four areas of vulnerability are identified, namely individualised political messaging; group dynamics and political polarisation; platform algorithms and self-radicalisation; and false news dissemination. Brief overviews are provided, in the following four sections, on each of these areas. Potential policy reactions are then discussed in the final section of this chapter.

2. INDIVIDUALISED POLITICAL MESSAGING

The ubiquitous use of multiple computer-based and/or mobile applications, by billions of users worldwide — and by almost every individual in advanced countries — is leading to the generation and collection of very large volumes of very granular data ("big data"). This includes, for example, purchases of goods and services; search engine queries; and emotional responses to a large array of online content, from news stories to popular memes, entertainment or leisure activities, and of course to commercial advertising and political campaigns. The average individual in a Western nation today has already voluntarily released thousands, if not millions, of data points into various software applications, almost always without any awareness as to how such data might be used and what insights might be gained from cutting-edge analysis.

Based on these developments, new ground has been broken in the academic field of psychometrics, and the corresponding applied field of *psychographics*. Recent analyses have revealed the close connection between individual preferences and behaviour, and private characteristics. As early as 2013, academics had demonstrated[123] that Facebook 'likes' could be used to automatically and (largely) accurately predict an individual's sexual orientation, ethnicity, religious and political views, personality traits, intelligence level, state of happiness, use of addictive substances, age, and gender. It is important to note that these results are not dependent on uncovering overt self-reporting of any of these characteristics. Rather, thanks to big data, psychometric research has revealed hitherto poorly understood correlations between overt online behaviour and intimate private information. These advances in psychometrics have revolutionised both marketing and political campaigning, based on improvements in *predictive analytics*, i.e. the use of data, statistical algorithms and machine learning techniques for purposes of prediction. A third key development is that differentiated political messages can now be delivered far more easily and cheaply down to the level of the individual voter through social media. Based on access to individual-level data, and to the deployment of new insights from psychographics and from predictive analytics, political messaging may be differentiated according to individual characteristics and personality traits in order to have the greatest psychological impact. In addition, each campaign ad can use a presentation format (e.g. colours, text size, choice of words, visual illustrations) whose emotional appeal has been optimised for its target audience, thanks to machine learning techniques[124]. The contemporary cutting-edge in political campaigning is thus reliant on a trinity of psychographics, predictive analytics, and individualised political messaging.

This new structure can be weaponised by hostile actors — leading to more effective campaigns of political subversion. At an October 2017 US Senate sub-committee hearing[125], it was revealed that Russian operatives had, for example, specifically targeted patriotic-minded adult Texans with a political advertisement purporting to be from a Texas-based organisation, and which contained false claims against then-candidate Hillary Clinton. Ads of this nature used Facebook's own targeting technology. The company delivered sophisticated and targeted political messages to audiences within the United States, in the context of a US election, in exchange for payment from an organisation under the ultimate control of the Kremlin. Compared to Cold War-era Soviet disinformation campaigns, it is striking how easily, quickly, and cheaply Russia was able to reach audiences in a Western country.

3. GROUP DYNAMICS AND POLITICAL POLARISATION

A key question is whether social media increases political polarisation. Overall societal or political polarisation is driven by many top-down and bottom-up factors, from the chosen attitudes and statements of politicians and the editorial lines of influential media sources to real-life socio-economic and societal shifts. *Group polarisation*, the phenomenon whereby joining a group of like-minded persons tends to entrench and sharpen pre-existing views, appears to be natural and was documented much before the advent of social media[126]. Seeking out information that conforms to pre-existing views (*selective exposure*), and finding such information more persuasive than contrary information (*confirmation bias*) are likewise not new phenomena[127]. The fact that group polarisation occurs also on social media, as shown in recent research[128], is thus no surprise. But it is not automatically obvious that social media would necessarily lead to greater polarisation for societies as a whole: individuals face a wide range of groups to choose from, including many moderate ones within which individuals would become entrenched moderates. If the choice and visibility of social media groups reflected the pre-existing or underlying distribution of opinion in society, social media might merely mirror that distribution.

However, if more extreme groups could benefit from undue advantages in the online world, then polarisation dynamics could be stronger than initial conditions in the offline world, and potentially lead to greater polarisation both online and offline. One angle of investigation is the rise — and partial mainstreaming — of anti-establishment and anti-liberal populism, in line with either far-right or far-left views, and often accompanied by sympathies or connections with the Kremlin. While the Great Recession of 2009 and its aftermath should be expected to have fuelled greater ‘demand’ for such views (without their Kremlin component), social media have not only facilitated the ‘supply’ of relevant content to receptive audiences but have also allowed such content to appear to be more popular than it is. Indeed, it has been shown that populist politicians and parties throughout the Western world enjoy considerably higher levels of support in the online world than they do at the ballot box. According to a Brookings Institution report[129], Germany’s AfD party has a Facebook following twice the size of that of the CDU party, although the latter obtained more than double the number of votes as the former in the 2017 election. Members of the European Parliament from the far-right have their tweets shared on average almost five times more than those from mainstream parties, and MEPs from the far-left account for 30% of the twitter followers of all MEPs, despite holding only 4% of seats. Overall, the (apparent) online popularity of extremist politicians seems to exceed their electoral popularity by a factor of at least four. While some of that gap could be explained by a willingness of some voters to express discontent only online, it is likely that fictitious online support is mostly to blame.

Fictitious online support is based on a combination of both home-grown and foreign *astroturfing*, i.e. the practice of masking the sponsors of a message or organization to make it appear as though it originates from grassroots participants. Online political astroturfing may combine both human agents (trolls) and automated agents (bots). It is not hard to see how online astroturfing may lead to distortions that could adversely affect the political process. Real-life swing voters could be swayed by the apparent size, popularity, and normalisation of extreme views and content. In addition, traditional media sources, and individual journalists, authors, and commentators, will typically be influenced by the number of likes and shares their works receive. If extreme content is artificially rewarded, this will create an (apparent) incentive to produce more of it, leading to (further) polarisation in traditional media, and to further polarisation of the electorate – both directly, as real-life voters receive more polarised content, and indirectly, as extreme networks and groups on social media are able to further legitimise their points of view by pointing to supporting content from traditional media, rather than only from fringe sources.

4. PLATFORM ALGORITHMS AND SELF-RADICALISATION

In 2017, 45% of Americans reported obtaining news stories from Facebook, 18% from YouTube, 11% from Twitter, and 7% from Instagram[130]. The news content on these platforms originates from third parties, notably from the web-sites of television, radio, and print media outlets. In addition, there is a large volume of user-generated content which is political in nature, regarding particular politicians and, perhaps more importantly, on current political, social and cultural issues.

Social media platforms aggregate and filter content according to user preferences. While much of that filtering can be traced back to conscious choices by users, e.g. choosing to 'follow' or 'friend' specific opinion leaders and self-selecting into specific groups, the algorithms used by the platforms generate individualised 'news feeds' as well as suggestions to follow or like or join additional opinion leaders or groups. These algorithms are partly based on machine learning techniques, and lead to filtering and selection criteria that are not transparently known. From the perspective of the platform operators, the goal is to retain the attention of users for as long as possible, given that longer attention time translates into greater exposure to advertising, and thus into greater revenues. Insights into human psychology, including users' individual personality traits, can be harnessed to maximise revenue-generation. This may incentivise platform operators to seek to generate obsessive, compulsive, or addictive emotional states — even beyond the platform operators' own explicit awareness.

An under-researched area of concern is the YouTube algorithm. As noted above, YouTube is a more important source of news than Twitter. Furthermore, surveys on sources of *news* likely underestimate the platform's true importance in indirectly shaping political perceptions through issues-based content that strongly correlates with political positioning (e.g. gender issues, climate change, immigration). The YouTube algorithm tends to automatically suggest content that goes in the same general direction as a viewer's interests, as expressed by the viewer's choices of videos. The algorithm is clearly oriented towards keeping users emotionally engaged, regardless of the factual accuracy of the content that is suggested to them. This is particularly problematic when it comes to politically charged issues, given that the platform contains a large volume of tendentious content, ranging quite seamlessly from subtly oriented and mostly true content to emotionally dark, alarming, or outright false content. This notably applies to "anti-establishment" views from the far-right or the far-left. As noted by some critics[131], this can result in leading some users down "hateful rabbit holes". Recent academic research[132] also notes the strong visibility of controversial content, while facing difficulties in elucidating how the algorithm works.

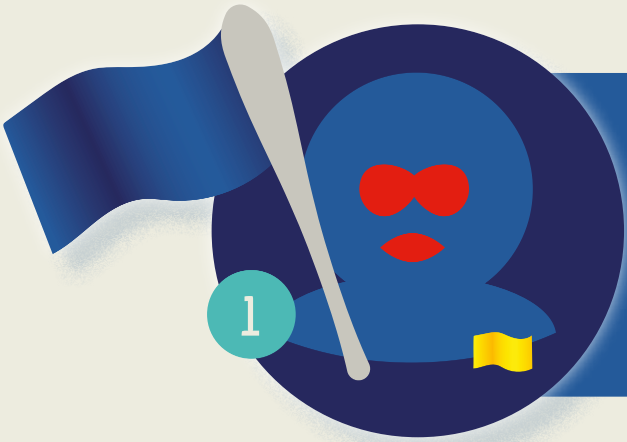
Risks to political stability arise from such algorithms for two main reasons. The first is self-radicalisation: users with merely a slight predisposition towards radical views are likely to consume far larger quantities of tendentious or false content than in the offline world. The second is that hostile actors may study how the algorithms promote certain types of content, and design their information operations accordingly.

5. FALSEHOOD DISSEMINATION DYNAMICS

The traditional tool-box of Soviet political subversion included disinformation operations. The latter involved the production of carefully fabricated falsehoods, and their injection into the information space of the adversary using a variety of channels and relays, ideally including ones that appear far removed from the originator. In some cases this included the production of forgeries, e.g. forged official letters to falsely attribute objectionable intentions to Western governments, or fake scientific research regarding the origins of the AIDS virus[133], or in simpler cases just the dissemination of a false rumour. Contemporary Russian disinformation is not essentially different, except that online technologies have greatly multiplied the speed and potential reach and depth of such operations.

Based on a large-scale analysis of 126,000 news stories over the 2006–2017 period on Twitter, a recent study[134] found that falsehood diffused significantly farther, faster, deeper, and more broadly than the truth. The researchers found that bots did not spread false news significantly more than true news over their particular sample, though one should bear in mind that the study does not distinguish between misinformation and disinformation. Other research and analyses make clear that bots are an important component of hostile disinformation campaigns. Most importantly, researchers found that human users are more likely to spread falsehood than the truth[135]. As with the other issues highlighted in this chapter, core vulnerabilities in the political information space stem from natural human psychological traits which can be exacerbated and exploited in online environments. The natural tendency of humans to spread false news, the group polarisation issues discussed previously, and the phenomena of confirmation bias and selective exposure, are sobering reminders that any official efforts to compete with falsehood on the basis of publishing the truth are bound to have only limited success.

The four public policy principles

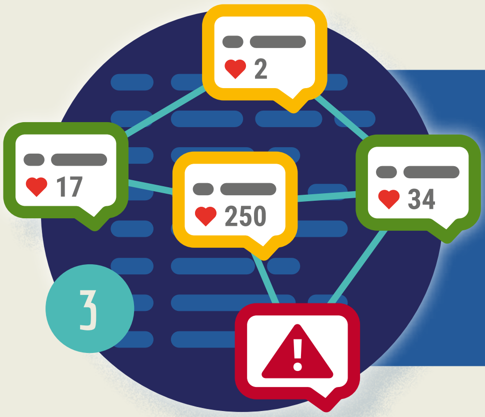


FALSE-FLAG INDIVIDUALISED POLITICAL MESSAGING:

Voters should know who is addressing political messages to them. False-flag messaging should be reduced as much as possible. Also, hostile foreign actors should not be permitted to promote any kind of political messaging in the context of domestic political campaigns.

POLITICAL ASTROTURFING OPERATIONS:

The fairness of the political process is endangered if malign actors are able to tip the scales in favour of any particular political actor or group of actors. Measures should be taken to prevent or block political astroturfing operations, or to render them ineffective.



CONTENT-SELECTION ALGORITHMS:

Without prejudice to principles of free enterprise and to the promotion of technical innovation, it should not be the case that content-selection algorithms generate individualised information spaces that entertain unduly extreme, delusional, obsessive, or paranoid mental states.

DISINFORMATION:

Liberal democracies must actively defend the integrity of their domestic political discourse from disinformation. Without prejudice to the freedom of expression and conscience of ordinary members of the public, the ability and capacity of hostile foreign actors to successfully carry out such operations should be degraded.

6. PROPOSED PUBLIC POLICY PRINCIPLES

This paper has highlighted four areas of vulnerability in contemporary liberal democratic systems. On the basis of these observations, four public policy principles are formulated below. A discussion then follows with selected examples of steps taken so far in selected jurisdictions. In many cases, the responses imply a need for a range of actions, including new legislation at national level and new (or strengthened) programmes of activities on the part of governmental and/or inter-governmental organisations. The active participation of major platforms is a necessary condition for the success of certain measures. In some cases, voluntary actions by platforms may be sufficient. In other cases, new legal obligations will prove necessary to ensure that the public interest is protected.

Implementing the four principles: some recent developments

False-flag individualised political messaging

Major platforms should systematically add clear labels onto paid political messaging. The latter should identify not only the formal name of the sponsoring person(s) or organisation(s) but, given the widespread phenomenon of front organisations, also the identity of the ultimate sponsor. If the ultimate sponsor is a foreign entity, paid political messaging should not be accepted to begin with. These considerations have been reflected in new legislative proposals in the United States and in France.

In the United States, a bill for an "Honest Ads Act" was introduced in the Senate in October 2017. The desired legal principles on clear labelling of political campaign messages, and on banning campaigning by foreign entities, already exist in US electoral law, namely in the Federal Election Campaign Act of 1971, but are not clearly applicable to the case of online platforms. The Honest Ads Act is thus a long overdue measure to ensure that political campaign rules that apply to radio, television, and print media also apply to online platforms. The Act would also obligate platforms to maintain a public file of all electioneering communications purchased by a person or group who spends more than $500 total on ads published on their platform. In April 2018, both Facebook and Twitter publicly stated their support for the bill, although media reports suggested there had been efforts by Facebook lobbyists to convince lawmakers to trust their platform with a purely voluntary approach[136]. The bill has yet to go through the legislative process.

In France, a proposal for a new law on the fight against false information[137] was submitted by members of President Macron's party in the National Assembly in March 2018. The law would notably amend France's electoral code, as well as its broadcasting law. At the time of writing, it remained to be seen what the final outcome would be after going through the legislative process[138]. With respect to false-flag political messaging, the French proposal foresees that platforms would be obligated to provide users with accurate, clear, and transparent information on the identity and nature of the persons or organisations (and on whose behalf they operate, if applicable) that have paid the platform to promote any content. Furthermore, platforms would have to make public the amounts received and the identity of the persons or organisations who paid for the promotion of informational content. In terms of substance, the part of the proposed French law that amends France's electoral law is thus highly similar to the proposed US legislation.

Political astroturfing operations

Ideally, astroturfing should be illegal as a matter of principle, and prevented from occurring. Activities such as open political debate (on social media platforms, on the websites of major media outlets), online petitions, and online public policy consultations, ought to be fair and transparent mechanisms that contribute genuinely-held views of real citizens into the political process. In practice, while astroturfing could easily be made illegal in principle, enforcement could prove challenging without the introduction of quite stringent identity checks. Assuming the latter route isn't pursued, certain steps could nevertheless be taken while retaining largely open avenues for voluntary expression.

First, public authorities should more forcefully accept that, whenever a policy issue is contested by hostile foreign actors, there is a risk that those actors will launch astroturfing operations. For policy-making processes that require consulting the public, governments and institutions such as the European Commission should make greater use of more reliable mechanisms, e.g. focus groups, polls based on random sampling, and statutory requirements or guidelines should be amended accordingly.

Second, for political campaigns, public authorities should work in partnership with industry and with independent researchers in order to improve the detection of fake online identities and of astroturfing campaigns, and to develop and test a range of options to reduce the impacts of such campaigns.

On the legislative side, the proposed new French law on the fight against false information includes an emergency mechanism for electoral campaign periods. Under this mechanism a designated judge could take any measures necessary, including shutting down websites, if state authorities detect a case where false information 'that would alter the fairness of the upcoming vote is *disseminated artificially and massively* to the public through an online platform'. The intention is to give a legal basis for stopping a pre-planned operation that would rely on networks of trolls and bots seeking to make a false story go viral. While such a provision would certainly be useful, long-term astroturfing campaigns would remain unaddressed.

Third, nations may choose to retaliate, as well as to use threats of future retaliation in order to deter the adversary. In response to Russian meddling in its 2016 election, the United States imposed sanctions[139] against the FSB, the GRU, and other Russian entities in December 2016. In March 2018, the US Treasury imposed additional sanctions[140], on named senior directors of the GRU for election-related cyber-attacks, and on named employees or associates of Russia's so-called Internet Research Agency, specifically due to the fact that they had "created and managed a vast number of fake online personas that posed as legitimate U.S. persons to include grassroots organizations, interest groups, and a state political party on social media [and] posted thousands of ads that reached millions of people online". Sanctions were thus brought about in response to both false-flag political messaging and political astroturfing.

Content-selection algorithms

While the algorithms of major platforms are proprietary in the sense of commercial law, there is a public interest case for some form of independent scrutiny. The best approach would be to designate authorised independent bodies to audit systemically important algorithms in order to mitigate risks relating to political polarisation, extremism, self-radicalisation, and ultimately to individual mental health and societal and political stability. This was first proposed in an earlier version of this paper[141]. The idea of auditing important algorithms without necessarily releasing them to the public was suggested more recently by author and mathematician Cathy O'Neil[142]. Building on these suggestions, states could draw inspiration from other cases of state regulatory bodies, with governance and financing arrangements that ensure independence from both government and industry, and with a legal obligation to respect the confidentiality of proprietary information. Such a body should have the power to instruct a major platform to modify algorithms it uses, and to demonstrate that changes have been implemented in such a way as to achieve designated outcomes. In the European context, it may be desirable given the cross-border nature of the phenomenon to directly seek the creation of a single EU-wide regulator under European law.

Disinformation

The countering of disinformation has generated the greatest response among the areas identified in this chapter, though activities so far (monitoring, flagging, debunking) have focused on reactive measures that, implicitly, accept the battlefield the way it is, rather than try to shape it (e.g. through new legislation). Many Western governments have created special inter-ministerial task-forces[143], typically involving their Interior Ministries or Justice Ministries as a priority, alongside other ministries (most often both Defence and Foreign Affairs), as well as dedicated centres in a smaller number of cases. Both the EU's External Action Service (EEAS) and NATO's Public Diplomacy Division have dedicated programmes and budgets to monitor and respond to disinformation. A group of NATO Allies also created the NATO Strategic Communications Centre of Excellence in Riga, which produces analysis and research, and provides expertise and training on countering hostile information activities by state and non-state actors.

Recent work at EU level — notably a major EU Joint Research Centre Technical Report[144], and a Communication (i.e. an official policy announcement) by the European Commission[145], both published in April 2018 — call for a series of measures to increase resilience in the face of disinformation. Both the report and the Communication call for greater transparency on the part of platforms, in order to address challenges such as false-flag political messaging. The European Commission's chosen approach so far is to encourage platforms to develop their own solutions, on the basis of a new, EU-wide Code of Practice on Disinformation, rather than to resort to new legislation. In addition, the European Commission wishes to develop an independent network of European fact-checkers: to better support debunking efforts; to develop positive incentives to foster quality journalism, including the development of initiatives to help media journalists to better handle instances of disinformation; and to encourage media literacy among the general public. A further area of interest which the European Commission proposes to support is the development of improved technologies, based on artificial intelligence, to identify, tag, and verify disinformation.

All of the measures above are well justified and ought to be pursued. One may however question whether they have quite the right intensity and depth of scope to counter the severity of the challenges that have emerged. For instance, while debunking efforts are necessary, available psychology and social psychology research strongly suggests that corrective stories may have rather limited impacts on large categories of voters. A sobering exercise in this respect is to contrast the viewership obtained by official debunking efforts, versus those obtained by hostile disinformation stories. For this reason, in some national cases, disinformation has been more effectively tackled by legal bans for certain information outlets — for instance in Latvia in April 2016, when the Rossiya RTR channel was banned for a 6-month period for broadcasting the views of a Russian politician who was inciting hatred and promoting military aggression.

In France, the proposed law on the fight against false information includes a strengthening of provisions to the existing broadcasting law in order to be able to revoke, or never grant, or temporarily suspend the broadcasting license of outlets that are controlled by a foreign state, *or under the influence of that state*, if the broadcaster harms the fundamental interests of the nation or takes part in a campaign of destabilisation of the nation's institutions, notably through the dissemination of false news.

07

Conclusion

Giorgio Bertolin

Throughout this publication, we have explored some of the ways in which technology can be exploited to influence our behaviour. In this brief section, we offer some recommendations.

Given current trends, it is likely that more and more personal data will be available online in the coming years. This issue is particularly delicate with regard to the personal data of people belonging to sensitive categories (such as servicemen/-women, government officials, and decision-makers). On the one hand, it will be necessary to devise ways to curb current trends for what concerns these categories, i.e. reducing the amount of information available on these individuals. On the other hand, it is equally important to consider how to mitigate the negative effects of data proliferation once data is already available and can be exploited by malicious actors. The research presented here has demonstrated that current standards must be improved to reduce the risks posed by personal data exploitation. Conducting experiments, such as the one described in this volume should be a staple component of tactical-level exercises and could significantly improve the awareness of our servicemen/-women.

The importance of visual material in social media analysis tends to be overlooked. We have outlined a possible methodology for the study of this type of content. Analysts should first ask themsleves how relevant visual material is for the topic they are monitoring. In most cases, adding visual material will considerably improve the range of vision of those exploring the taxonomy of narratives in the online environment.

In the short-to-mid term, It is unlikely that any NATO country will find itself in the same set of conditions that led Ukraine to enforce a ban on Russia-based social media platforms. However, the lessons that can be drawn from the study transcend the peculiarities of the case. It is currently impossible to unequivocally evaluate the ban as either ‘effective’ or ‘non-effective’, as the unanticipated side effects that have emerged may well outweigh those elements that are seen as effective. A ban enforced in a similar scenario would likely bear similar results, i.e. loss of popularity for the affected platform together with the radicalisation of discourse among those users who circumvent the norm. Since the current results are inconclusive, we do not recommend that any other country to follow in Ukraine’s footsteps at this time.

We have shown how information activities carried out online follow different scripts depending on whether they are intended for external or internal audiences, and how real-world events constitute the catalyst for the dissemination of selected narratives online. Our societies are targeted by narratives that aim at exacerbating existing divisions. Directly countering these narratives would only spread their message further. However, it is possible to considerably reduce their appeal by spreading positive, inclusive, forward-looking narratives. This way, divisive messages will not be able to inhabit the narrative void and infect the social organism.

ENDNOTES

1 David Reinsel, John Gantz, and John Rydning, *Data Age 2025: The Evolution of Data to Life-Critical* (International Data Corporation, 2017). Accessed September 13, 2018.

2 Khoso, Mikal, 'How Much Data Is Produced Every Day?', Level Blog (blog), Accessed May 13, 2016.

3 Cambridge Analytica website. Accessed September 13, 2018.

4 Wagner, Kurt, 'Here's How Facebook Allowed Cambridge Analytica to Get Data for 50 Million Users', *Recode*, Accessed March 17, 2018.

5 *Washington Post*, Mark Zuckerberg Testifies on Capitol Hill (Full Senate Hearing video), 2018. Accessed September 13, 2018.

6 Privacy International website, 'Data and Elections | Privacy International'. Accessed September 13, 2018.

7 Privacy International website, 'What Zuckerberg Forgot To Mention | Privacy International'. Accessed September 13, 2018.

8 Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez, 'Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones', 2011 15th Annual International Symposium on Wearable Computers (San Francisco, CA, USA: IEEE, 2011): 29–36.

9 Jagdish Prasad Achara, Gergely Acs, and Claude Castelluccia, 'On the Unicity of Smartphone Applications' Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society — WPES 2015 (Denver, Colorado, USA: ACM Press, 2015): 27–36.

10 Driss Choujaa and Naranker Dulay, 'Predicting Human Behaviour from Selected Mobile Phone Data Points', Proceedings of the 12th ACM International Conference on Ubiquitous Computing — Ubicomp 2010 (Copenhagen, Denmark: ACM Press, 2010): 105.

11 Privacy International website, 'Examples of Data Points Used In Profiling'. Accessed September 13, 2018.

12 Deceptive elements, intention to do harm, disruptive, interference with domestic democratic processes (DIDI).

13 Pamment, James, Howard Nothhaft, Henrik Agardh-Twetman, and Alicia Fjällhed. "Countering Information Influence Activities: The State of the Art," Lund University, 2018.

14 Companies include Spokeo, PeekYou, PeopleSmart, Pipl, and others.

15 Companies include Datalogix, Experian, Equifax, and others.

16 Companies include ID Analytics, Biznode, UC, and others.

17 Giridhari Venkatadri, Athanasios Andreou, Yabing Liu, Alan Mislove, Krishna P. Gummadi, Patrick Loiseau, and Oana Goga, 'Privacy Risks with Facebook's PII-Based Targeting: Auditing a Data Broker's Advertising Interface', 2018 IEEE Symposium on Security and Privacy (SP), (San Francisco, CA: IEEE, 2018): 89–107.

18 Ibid.

19 Facebook Newroom website, 'A New Level of Transparency for Ads and Pages | Facebook Newsroom'. Accessed September 13, 2018.

20 Dipayan Ghosh, 'Facebook Is Changing How Marketers Can Target Ads. What Does That Mean for Data Brokers?' *Harvard Business Review*, Accessed April 9, 2018.

21 A group of people that assume an adversial role during a military exercise.

22 Our experimentation focused on assessing the availability of data rather than actual data collection. Further safeguards were implemented by strict controls of all individuals with access to the experiment, regularly deleting data which might have accidentally been collected by automated processes such as web caches, and implementing a strict procedure in order to ensure that no personal data was stored after the end of the experiment.

23 United States Department of Justice website, 'Grand Jury Indicts Thirteen Russian Individuals and Three Russian Companies for Scheme to Interfere in the United States Political System', 16 February 2018.

24 Pretending to be another person for the purpose of fraud.

25 Pages designed to lure or attract a certain group of people.

26 Use of deception to manipulate individuals into providing confidential or personal information.

27 Search engines designed to find information about physicaly persons. Some examples of people search engines include Spokeo, PIPL etc.

28 The page was reported.

29 Note, users did not have to accept our friend request for FB to start suggesting similar friends.

30 Mark Zuckerberg, 'Preparing for Elections', Facebook. Accessed 13 September 2018.

31 Federal Trade Commission, *Data Brokers: A Call For Transparency and Accountability*, (Federal Trade Commission, 27 May 2014).

32 DAESH Information Campaign and its Influence, NATO Strategic Communications Centre of Excellence, Riga, 2016

33 "Mary Meeker'S 2016 Internet Trends Report: All The Slides, Plus Highlights". 2018. Quartz.

34 Picture made with and sent via Snapchat, a multimedia messaging application.

35 Taking a picture of something and publishing it on Instagram, a photo sharing platform.

36 Data item that carries information in an unstructured format and is hard to quantify.

37 Although these tools have 'social media' in their names, they are also used to crawl the wider web, i.e. news sites, blogs, forums, etc.

38 Machine learning algorithms are algorithms that, given annotated data, progressively improve their performance at a specific task, e.g. classifying images or predicting trends. A model is an algortihm that has been trained using data and validated as having learned to preform the task in question.

39 Some current state-of-the-art algorithms include You Only Look Once (YOLO), Single-Shot Detector (SSD), and Regional Convolutional Neural Network (RCNN).

40 "Review Of Deep Learning Algorithms For Object Detection". 2018. Medium.

41 ImageNet with 500 000 images and 200 categories, COCO with 120 000 images, PASCAL VOC with 10 000 images and 20 categories.

42 Bermeitinger, Bernhard & Radisch, Erik & Howanitz, Gernot. (2018). Contextualizing Bandera: Ein Distant Watching-Ansatz

43 An API endpoint most often is a URL to a service or server. Each endpoint is the location from which APIs can access the resources they need to carry out their function, e. g., find and describe faces in an image.

44 Real-world objects, such as people, locations, organisations, products, etc., which can be denoted with a category. Named entities are extracted from text as part of computer-based text analysis.

45 "Lithuania Looking For Source Of False Accusation Of Rape By German..." 2018, Reuters, U.S.

46 Data extraction from websites.

47 A method for unsupervised machine learning developed by Teuvo Kohonen in 1980s, also known as Kohonen maps. SOMs build a two-dimensional map of multivariate data by assembling data items in nodes of similar variables.

48 Graphical representation of data that uses a system of color-coding to represent different values.

49 14–20 September 2017.

50 International news agency that reported on NATO intercepting Russian jets near Estonian airspace.

51 Soviet fighter planes and interceptor aircraft used during World War II.

52 This refers to FA-18 hornets, fighter jets used in NATO's Baltic air-policing mission.

53 Initially it was conducted for all datasets, but due to the low volume of data for the 'NATO in Poland and the Baltics' datasets, the SOMs yielded no meaningful maps.

54 NATO military operation against the Federal Republic of Yugoslavia during the Kosovo War in 1999.

55 VK-users whose posts were identified as ideological by the study's clustering algorithm.

56 These statistics include non-unique visits from desktops. According to the SimilarWeb website analysis tool, in most cases these data include visitors from Ukraine who accessed the banned sites via VPN.

57 Google Trends data; different languages accounted trends.

58 Retrieved from Similarweb on 18 March 2017

59 Retrieved from Similarweb on 14 August 2018

60 Retrieved from Alexa on 14 August 2018

61 852 690 profiles

62 Self-reported location: 3 436 profiles changed their location from Ukraine to Russia, 276 moved to the USA, 266 to Poland, 96 to Belarus, 93 to Germany, 73 to Japan, 68 to Kazakhstan, and 824 to other countries, and 6,109 no longer specify a location.

63 Users who wrote or reposted more than one post during the period studied

64 List of areas

65 Some users located in Crimea don't recognize it as part of Ukraine, but are unable to choose Russia from the drop-down list provided by VK for this territory. If users didn't choose 'Ukraine' as the location of their profile, they were not added to the sample

66 Despite the fact that the Ukrainian government has banned VK, it can't be blocked there because telecommunications have been 'nationalized' and are not under the control of the Ukrainian government

67 General statistics before the ban: min = 60,067 posts , max = 101,175 posts , mean = 80,388 posts

68 General statistics after the first exodus: min = 26,369 posts, max = 42,832 posts, mean = 37,688 posts

69 General statistics after the second exodus: min = 23,561 posts, max = 30,009 posts , mean = 27,191 posts

70 Despite the fact that the official date of the ban was 15 May 2017, Internet providers were not able to ban sites immediately. The ban was executed over the following weeks

71 Number of users per capita

72 Number of users per capita

73 Profiles for which the specified year of birth was earlier than 1940 were excluded from the set

74 Mobile applications available from Google Play and AppStore were used to access some of the banned domains

75 Google Trends data; different languages accounted; Telegram data is given on a larger scale

76 An active user is a user who generated at least one post before the ban

77 We regarded a user as having left the social network if they stopped posting

78 299 370 posts before the ban and 289 733 after the ban

79 Posts containing 87 stop words associated with ‘VK games’, ‘movies’, ‘gifts’, and ‘recipes’ were considered spam. We also used an artificial neural network to classify posts and label them as spam if they were related to such topics as ‘movies’, ‘shopping’, ‘gifts’, ‘food’, or ‘music’; topics such as ‘religion’, ‘vata’,and ‘vatnik’ were retained in the sample

80 An ideological user is a user whose posts contain ideologically-tinged words or word patterns. Ideologically-tinged traffic was identified by a clustering algorithm applied to random samples of ~ 600 000 posts.

81 There were 12,186 ideological posts before the ban and 14, 888 posts after the ban, or 4.07% and 5.14% of all posts in the first and second samples respectively

82 19 974 posts during the studied period (or 25.8 posts per day)

83 Here, a connection means having three or more mutual friends

84 19 988 posts during the period studied (or 41 posts per day before the ban and 11.2 posts per day after the ban)

85 Nadiya Savchenko

86 We considered profiles as having left VK if they generated 0 posts after the ban — 103 ideological users (or 22%) ceased posting after 5 June 2017; 98 profiles were from the GCA and five profiles were from the NGCA

87 22 449 groups

88 From 21 709 groups, the rest were either blocked, deleted, frozen, or closed to non-subscribers

89 Ideological groups

90 11 186 posts per day from 620 groups on average; a maximum of 14 663 posts

91 12 170 before and 10 135 after the first exodus of users leaving VK

92 The sample contains an average of 702 reposts per day by Ukrainian users before the ban 5 June 2017,and an average of 280 reposts per day afterwards

93 Views of the posts are counted for all VK users with no regard to the country, since in the most cases ideological groups on VK have a cross-national list of subscribers, thus there is an organic limit to the influence of the Ukrainian ban. These groups are associated with Ukraine because Ukrainian users have reposted their content significant number of times

94 67 347 428 before and 54 274 690 after the first user exodus

95 BBC: Russia fire: Children killed in Kemerovo shopping centre blaze

96 Pavel Durov VK page post, Page 1 of the order archived, Page 2 of the order archived

97 291 946 active users from the sample before the ban vs 92 622 current users. These values correspond to the results of other studies

98 One post per three days vs one post per four days

99 Born in 1991 vs born in 1987

100 Born in 1991 vs born in 1987

101 Giles, Keir (2017) Countering Russian Information Operations in the Age of Social Media, Council on Foreign Relations

102 Linvill DL, Warren PL (2018) Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building.

103 Russian Troll Factory tweets

104 Reddit’s 2017 transparency report and suspect account findings, Reddit Suspicious Accounts Release

105 Data Set: Removed Facebook Pages: Engagement Metrics and Posts, Social Media Advertisements, U.S. House of Representatives, Permanent Select Committee on Intelligence

106 Data Set: Top Instagram Posts By Likes and Comments Over Time

107 Exposing Russia’s Effort to Sow Discord Online: The Internet Research Agency and Advertisements

108 Instagram, Meme Seeding, and the Truth about Facebook Manipulation, Pt. 1

109 In particular, Jonathan Albright, Tow Centre for Digital Journalism, and the team and FiveThirtyEight.

110 Exposing Russia’s Effort to Sow Discord Online: The Internet Research Agency and Advertisements

111 Tom Postmes, Building or Breaching Social Boundries? SIDE Effects of Computer-mediated Communication, *Communication Research* 25(6) (1998): 689–715.

112 C. McGarty, *Categorization in Social Psychology*, (London, Thousand Oaks, New Delhi: Sage Publications, 1999).

113 C. Bail C et al., ‘Exposure to opposing views can increase political polarization: evidence from a large-scale field experiment on social media’, *Procedings of the National Academy of Sciences of the United States of America* (2018):1–6.

114 K. Müller K and C. Schwarz C, Fanning the Flames of Hate: Social Media and Hate Crime, 2017. *Ssrn*. doi:10.2139/ssrn.3082

115 Rolf Fredheim, Robotrolling 2018/1, NATO Strategic Communications Centre of Excellence, 2018.

116 Google Project Jigsaw, ‘Perspective’. Accessed: 23 March 2018.

117 E. Wulczyn, N. Thain, and L. Dixon, ‘Ex Machina: Personal Attacks Seen at Scale’2016, pp. 1–9.

118 H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, ‘Deceiving Google’s Perspective API Built for Detecting Toxic Comments’, 2017.

119 . Chu, K. Jue, and M. Wang, 'Comment Abuse Classification with Deep Learning', 2017, pp. 1–8.

120 The views in this chapter are solely those of its author, and do not necessarily represent those of NATO or of Allied governments. The author wishes to thank Armand De Mets, Vineta Mēkone, Jānis Sārts, Ulf Ehlert, Laura Brent, Christian Liflander, Chelsey Slack, Neil Robinson, Holly Vare, and James Reynolds-Brown for comments on earlier versions of the text. A stand-alone variant of this chapter was published as: Edward Hunter Christie, 'Political Subversion in the Age of Social Media', Policy Brief, Wilfried Martens Centre for European Studies, October 2018.

121 This definition was reported by Ladislav Bittman, a Cold War-era defector from the former Czechoslovak intelligence service, in 1984. Source: Soviet Active Measures, 1984 [film]. USA: US Information Agency.

122 J.-B. Jeangène Vilmer, A. Escorcia, M. Guillaume, J. Herrera, Les Manipulations de l'information : un défi pour nos démocraties, rapport du Centre d'analyse, de prévision et de stratégie (CAPS) du ministère de l'Europe et des Affaires étrangères et de l'Institut de recherche stratégique de l'École militaire (IRSEM) du ministère des Armées, Paris, August 2018, page 50.

123 Michal Kosinski, David Stillwell and Thore Graepel, 'Private traits and attributes are predictable from digital records of human behavior', *Proceedings of the National Academy of Sciences*, April 2013, 110 (15), 5802-5805.

124 This is notably based on a method called A/B testing, whereby small variations in online advertising materials (version A versus version B) are selectively shown to large numbers of otherwise similar potential customers to choose the most effective version. The process is repeated multiple times with successive variations, leading to particularly effective materials.

125 US Senate sub-committee hearing on "Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions", 31 October 2017.

126 See e.g. David G. Meyers and Helmut Lamm, 'The Group Polarization Phenomenon', *Psychological Bulletin*, 1976, Vol. 83, No. 4, pp. 602-627.

127 See e.g. Charles G. Lord, Lee Ross, and Mark R. Lepper, 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence', *Journal of Personality and Social Psychology*, 1979, Vol. 37, No. 11, pp. 2098-2109.

128 See e.g. Michela Del Vicario, Gianni Vivaldo et al., 'Echo Chambers: Emotional Contagion and Group Polarization on Facebook', *Nature*, Scientific Reports, 6:37825.

129 Clara Hendrickson and William A. Galston, 'Why are populists winning online? Social media reinforces their anti-establishment message', Brookings Institution, 28 April 2017.

130 Elisa Shearer and Jeffrey Gottfried, 'News Use Across Social Media Platforms 2017', Pew Research Center.

131 Paul Lewis, '"Fiction is outperforming reality': how YouTube's algorithm distorts truth', *The Guardian*, 2 February 2018.

132 See e.g. Bernhard Rieder, Ariadna Matamoros-Fernandez and Oscar Coromina, 'From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results', *Convergence: The International Journal of Research into New Media Technologies*, 2018, Vol. 24 (1), pp. 50-68.

133 See e.g. Thomas Boghardt, 'Operation INFEKTION: Soviet Bloc Intelligence and Its AIDS Disinformation Campaign', *Studies in Intelligence*, Vol. 53, No. 4 (December 2009).

134 Soroush Vosoughi, Deb Roy, and Sinan Aral, 'The spread of true and false news online', *Science*, Vol. 459, pp. 1146-1151 (2018).

135 Ibid.

136 Heather Timmons and Hanna Kozlowska, 'Facebook's quiet battle to kill the first transparency law for online political ads', *Quartz*, 22 March 2018.

137 The proposal can be accessed at: http://www.assemblee-nationale.fr/15/propositions/pion0799.asp

138 The National Assembly (the lower chamber) adopted its first reading position on 3 July 2018, essentially supporting the original text with relatively minor amendments. A setback occurred, as the Senate rejected the text outright on 26 July 2018, rather than produce a rival first reading position. However, France's legislative process gives the last word to the lower chamber when no consensus can be found between the two chambers.

139 Executive Order 13757 of December 28, 2016, Federal Register Vol. 82, No. 1, Tuesday, January 3, 2017.

140 US Treasury Department, 'Treasury Sanctions Russian Cyber Actors for Interference with the 2016 U.S. Elections and Malicious Cyber Attacks', Press Release, March 15, 2018.

141 Edward Hunter Christie, 'Artificial Intelligence in Tomorrow's Political Information Space', paper presented at the NATO STO Experts Meeting on Big Data and Artificial Intelligence for Military Decision Making, Bordeaux, 30 May — 1 June 2018, DOI: 10.14339/STO-MP-IST-160-PT-3-PDF

142 Cathy O'Neil, 'Audit the algorithms that are ruling our lives', Opinion, *The Financial Times*, 30 July 2018.

143 For a list of national examples, see J.-B. Jeangène Vilmer, op. cit., p. 119.

144 Bertin Martens, Luis Aguiar, Estrella Gomez-Herrera and Frank Mueller-Langer, 'The digital transformation of news media and the rise of disinformation and fake news — An economic perspective', *Digital Economy Working Paper* 2018-02; JRC Technical Reports.

145 European Commission, 'Tackling online disinformation: a European Approach', COM(2018) 236 final, 26 April 2018.

Prepared and published by the

NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org