Finding and Characterizing Information Warfare Campaigns

David Beskow

CMU-ISR-20-107

March 2020

Institute in Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley (Chair)
Douglas Sicker
Yulia Tsvetkov
Matthew Dabkowski (United States Military Academy)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Societal Computing.

The research reported in this thesis has been supported in part by the Office of Naval Research Award Numbers N00014182106, N0001418-1-2108, N00014-18-SB001, N000141812108, N000140811186, Office of Naval Research Minerva Awards N00014-13-1-0835/N00014-16-1-2324, the Knight Foundation and the Center for Computational Analysis of Social and Organization Systems (CASOS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

Keywords: social cyber security, information warfare, machine learning, social network analysis, recommender engines, bots, internet memes

Abstract

Today the borderless internet is used by state and non-state actors to manipulate information and societies in ways that were unheard of 50 years ago. Malicious actors can rapidly conduct information maneuvers with little cost at unprecedented scales to achieve far reaching consequences across the internet. They do this by exploiting features of the various social media platforms and the way humans naturally understand what they read and hear. These cyber-mediated threats to open and democratic societies have led to an emerging discipline known as social cybersecurity.

While various aspects of these campaigns have been explored, little research has focused on the campaign level of engagement. Our research seeks to answer the question: How can information warfare campaigns be identified and characterized quickly? Our goal is to 1) Improve understanding of information operations, and 2) Develop techniques to rapidly identify key factors such as *bots* and *memes*.

To accomplish this, I present the strategic context of the information warfare that we see today, and identify and define information warfare *forms of maneuver*. I develop various supervised and unsupervised methods to identify bots at four different data granularities. I present a deep learning model to classify memes as well as study the evolution of memes within a conversation. I present a template for understanding the major components of an information campaign and develop automatic ways to populate this template for a specific event. Finally, we present a Bot, Cyborg, and Troll Field Guide to help analysts and the general population understand these entities.

Acknowledgments

Support for this work came in part from the funding agencies and grants listed below:

- Office of Naval Research Grant N00014182106 Group Polarization in Social Media
- Office of Naval Research Grant N0001418-1-2108
- Office of Naval Research N00014-18-SB001
- Office of Naval Research N000141812108
- Office of Naval Research Minerva Awards N00014-13-1-0835/N00014-16-1-2324
- Office of Naval Research Multidisciplinary University Research Initiative Award N000140811186
- Army Research Laboratory Award W911NF1610049
- Defense Threat Reductions Agency Award HDTRA11010102

I want to thank West Point and the Army Advanced Civil Schooling Program for the opportunity to pursue a PhD at Carnegie Mellon University. I am grateful for travel funding provided by the Operations Research Center at West Point and the Omar N. Bradley Officer Research Fellowships in Mathematics. The Computational Analysis of Social and Organizational Systems (CASOS) Center provided funding, support, and collaboration, for which I am also grateful.

I want to especially thank my advisor, Dr. Kathleen Carley, for her hours of advice and mentorship, and the rest of my committee for their patience and timely advice. I'd like to recognize my fellow researchers at the Institute of Software Research, especially Matthew Babcock, Sumeet Kumar, Binxuan Huang, Iain Cruickshank, Ramon Villa Cox and Thomas Magelinski for your help, advice, and encouragement over the past few years.

Finally, I would like to thank my family for their selfless efforts to support my studies and research. I'd like to thank God for the capacity to think and reason. Finally and most importantly, I'd like to thank my Heavenly Father for His Son Jesus Christ, who came so we may "*...have life and have it abundantly.*" John 10:10b.

Contents

List of Figures

List of Tables

Chapter 1

Strategic Context

> "...a new-generation war will be dominated by information and psychological warfare that will seek to achieve superior control of troops and weapons and to depress opponents' armed forces personnel and population morally and psychologically. In the ongoing revolution in information technologies, information and psychological warfare will largely lay the groundwork for victory." [62]
>
> \- Russian *Military Thought*, 2013

> "Russia is waging the most amazing information warfare blitzkrieg we have ever seen in the history of information warfare."
>
> –Gen Philip Breedlove, NATO Wales summit, September 2014

1.1 Introduction

Social cyber security is an emerging subdomain of national security that will affect all levels of future warfare, both conventional and unconventional, with strategic consequences. Social cybersecurity "is an emerging scientific area focused on the science to characterize, understand, and forecast cyber-mediated changes in human behavior, social, cultural and political outcomes, and to build the cyber-infrastructure needed for society to persist in its essential character in a cyber-mediated information environment under changing conditions, actual or imminent social cyber-threats" [56]. Technology today is enabling both state and non-state actors to manipulate the global marketplace of beliefs and ideas at the speed of algorithms, and this is changing the battlefield at all levels of war.

While recently viewed through the lens of "hybrid" warfare, information warfare is becoming an end unto itself. Dmitry Kiselev, coordinator of the Russian state agency for international news, states that "information wars are...the main type of war" [259]. Information is used to strengthen your narrative while attacking, disrupting, distorting and dividing the society, culture, and values of other competing states and organizations. By weakening trust in national institutions, consensus on national values, and commitment to those values across the international community, an actor can win the next war before it has even begun. In fact, reflecting the change from periodic conflict to continual competition [232], Senior leaders in the Russian General Staff have claimed that "Wars are not declared but have already begun" [106].

Information is strengthening its position within the elements of national power. Strategy is often viewed through the Elements of National Power: Diplomatic, Information, Military, and Economic (DIME). Technology now allows state and non-state actors to extend their power in the *Information* domain at a scale and complexity long thought impossible. If left unchecked, this emerging 'information blitzkrieg' will have strategic effects on par with the physical blitzkrieg unleashed at the outset of World War II.

War is ultimately a human endeavor. The ways that technology transform humanity is at the heart of emerging trends. Human interaction has dramatically increased due to social media, and access to information and ideas has also dramatically increased. Social media allows rapid mobilization of masses around ideas. This technological trend offers opportunity to those who are adept at wielding informational power. Because of this, information may become the preeminent commodity and decisive factor in future conflict [117].

While technical in nature, social cyber security differs from traditional cyber security. Traditional cyber security involves humans using technology to 'hack' technology. The target is information systems. Social cyber security involves humans using technology to 'hack' other humans. The target is humans and the society that binds them. This twist on the traditional cyber paradigm is sometimes referred to as 'cognitive hacking.' While leveraging the cyber medium for mass delivery, this emerging information warfare leverages advances in targeted (or micro) marketing, psychology and persuasion, policy gaps at and between private and government institutions, and understanding of the social sciences to deploy coordinated information operations with strategic effect.

Social cyber security is inherently multi-disciplinary computational social science. "Emerging theories blend political science, sociology, communication science, organization science, marketing, linguistics, anthropology, forensics, decision science, and social psychology" [56]. Many researchers in this field are leveraging computational social science tools such as network analysis, spatial analysis, semantic analysis, and machine learning. These are applied at multiple levels, from the individual through the conversation level to the larger community level.

This chapter will introduce and define this emerging discipline, briefly discuss its history and the socio-technological changes that enable it, and finally discuss current and emerging social cyber security 'forms of maneuver'. Throughout this process we will elaborate on the similarities and differences between social cyber security and traditional cyber operations. Parts of this chapter were published in [34] and [35].

1.2 Understanding Information

> "Information is the Resolution of Uncertainty"
>
> –Claude E. Shannon, 1948

Before we define disinformation and misinformation, we first must develop and refine our notion of *information*. Following the massive increase in communications and data used during World War II, several individuals began to develop the theory of information. Norbert Wiener clarified that "information is neither matter, nor energy" [252]. Claude Shannon, a brilliant mathematician at Massachusetts Institute of Technology who worked on anti-aircraft algorithms during the Second World War, believed that information was "the most mathematical of the engineering sciences." He eventually refined his thesis to define information as the "resolution of uncertainty" [210]. For Shannon (see Figure 1.1), the easiest way to envision this resolution of uncertainty is with the flip of a coin. He envisioned that all information could be represented by a coin flip, or a binary category of heads and tails. The notion of the bit was born, and the computer and the entire digital age was built on this theory, launching what is known today as the "Age of Information" [156].

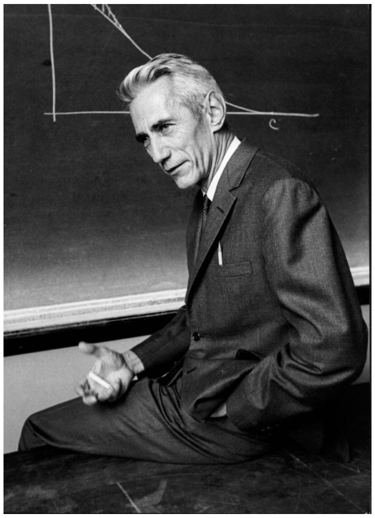


Figure 1.1: Claude Shannon, Father of the Bit and the Information Age

We intuitively understand that data is underlying information, but are data and information one and the same thing? A humorous quote among statisticians is "In God we Trust, all others must bring data" (attribution disputed). This quote, while humorous in intent, implies that data, like information, is an essential resource for "resolving uncertainty." The Data → Information → Knowledge → Wisdom Paradigm helps explain a bit of this difference. This paradigm is generally attributed both to a 1934 poem by T.S. Eliot entitled *The Rock* [88] and a 1989 article by Arkoff [5], and has had multiple additions and modification by various authors and organizations (for example, the US Army replaced/defined "Wisdom" with "Shared Understanding". While it does have some notable critics [269], it is nonetheless helpful for our endeavors. For a full explanation of the DIKW hierarchy, see Rowley's work [199] in addition to the original by Arkoff. In the DIKW hierarchy, *data* is an observation (made by a human or sensor) that is often recorded on computers using Shannon's bits. *Information* is often descriptive, and answers the basic questions of who, what, when, where, and how. *Knowledge* makes the information usable by humans by converting it to instructions. These instructions can be gained through transmission or through experience. Intelligence increases efficiency, while *wisdom* increases effectiveness, often adding values judgements (moral and ethical judgement). Rowley succinctly summarizes this by saying the data is simply bits, information provides the *what*, knowledge provides *how*, and wisdom provides *why* [199]. Having defined information as the "resolution of uncertainty" and expanding on this definition with the DIKW hierarchy, let's examine distortions of information.

1.2.1 Terms used for Information Manipulation

If information allows a person to "resolve uncertainty" in the search of truth, then manipulating information can cause that same person to arrive at an alternate truth. Today, nation states as well as various domestic and international actors manipulate data and information in order to sway the actions (knowledge) and beliefs (wisdom) of target audiences. This manipulation is undertaken to achieve strategic ends. Any change of facts (information) is considered *misinformation*, and includes both intentional and accidental manipulation. Accidental manipulation can happen early in reporting around an emerging event when well-meaning news and government agencies have not checked and verified facts (information). Disinformation is defined as misinformation that is intentional and harmful manipulation of data and information in order to change beliefs and both individual and collective action.

When information is used to influence a population, it is often called propaganda. Propaganda, while manipulative, is not necessarily false. Propaganda is simply using information, both fact and fiction, to influence. In this respect, not all propaganda is disinformation. In recent political dialogue, however, the lines of separation between disinformation and propaganda are blurring, and disinformation is becoming synonymous with propaganda, defining any manipulation of information, not just those that use lies [72].

Note that propaganda is often separated into three categories: *black*, *grey*, and *white* propaganda. *Black* propaganda is designed to appear as if it was created by the organization it aims to discredit. *Grey* propaganda attempts to hide the source of the propaganda, and *white* propaganda does not attempt to obfuscate its source, and is sometimes called *overt* propaganda.

1.2.2 Defining Information Operations

Defining information operations is the toughest definition to nail down. Social media companies seem to equate information operations with fake news, meaning that if all your facts are together, you're not conducting information operations. This also means that if you tell a social media company that your organization is conducting information operations, they believe that you are inherently spreading lies [51]. The Navy and Air Force focus more on the technical flow of information, and therefore information operations is primarily traditional cyber and electronic warfare. The Army and Marines focus more on the human side, and therefore information operations is more of a concerted effort to influence a target audience to change their beliefs or behavior [51]. Additionally, the US military view of information operations has primarily focused on tactical IO after the onset of hostilities. In fact, one expert has said that the West's view of "IW is almost by definition countercommand and control warfare" [44]. Russia and other nations, on the other hand, view it as an ongoing activity (during peacetime and war) [107]. Additionally, Russia does not distinguish between cyber and information warfare. They do not have a distinct cyber division in organization and thinking as the West (US/NATO) do [107], which allows them to more easily synchronize cyber and non-cyber information warfare.

Within US military doctrine, the most recent Joint Publication on Information Operation military doctrine breaks the information environment down into the cognitive dimension (human-centric), the information dimension (data-centric), and the physical dimension (real and tangible) (see Figure 7.1 in Chapter 7) [221]. It defines *information operations* as “the integrated employment, during military operations, of information-related capabilities in concert with other lines of operation to influence, disrupt, corrupt, or usurp the decision-making of adversaries and potential adversaries while protecting our own” [221]. Information-related capabilities are any “tool, technique, or activity employed within a dimension of the information environment that can be used to create effects and operationally desirable conditions” [221]. The manner with which information operations are used is further defined as

> “The first step in achieving an end(s) through use of the information-influence relational framework is to identify the target audience. Once the target audience has been identified, it will be necessary to develop an understanding of how that target audience perceives its environment, to include analysis of target audience rules, norms, and beliefs. Once this analysis is complete, the application of means available to achieve the desired end(s) must be evaluated. Such means may include (but are not limited to) diplomatic, informational, military, or economic actions, as well as academic, commercial, religious, or ethnic pronouncements. When the specific means or combinations of means are determined, the next step is to identify the specific ways to create a desired effect. Influencing the behavior of target audiences requires producing effects in ways that modify rules, norms, or beliefs. Effects can be created by means (e.g., governmental, academic, cultural, and private enterprise) using specific ways (i.e., information related capabilities) to affect how the target audiences collect, process, perceive, disseminate, and act (or do not act) on information. Upon deciding to persuade or coerce a target audience, the commander must then determine what information related capabilities it can apply to individuals, organizations, or systems in order to produce a desired effect(s). As stated, information related capabilities can be capabilities, techniques, or activities, but they do not necessarily have to be technology-based. Additionally, it is important to focus on the fact that information related capabilities may come from a wide variety of sources. Therefore, in IO, it is not the ownership of the capabilities and techniques that is important, but rather their integrated application in order to achieve a joint force commander’s end state. ” [221]

The increased importance of information in strategic competition is further highlighted by the US Department of Defense when the Chairman of the Joint Chiefs of Staff issued a change to Joint Publication (JP) 1, “Doctrine for the Armed Forces of the United States” introducing information as a new joint function [117]. In US Joint Doctrine “Joint functions represent related capabilities and activities placed into basic groups to help commanders synchronize, integrate, and direct operations” [117]. Adding *information* to the original 6 joint functions (Command and Control (C2), Intelligence, Fires, Movement and Maneuver, Protection, and Sustainment) shows a doctrinal re-prioritization of information in the modern strategic context. This addition acknowledges that, while conflict, violence, and war endure, the methods through which nations pursue political goals are evolving [205].

> The information function encompasses the management and application of information and its deliberate integration with other joint functions to influence relevant actor

> perceptions, behavior, action or inaction, and human and automated decision making. The information function helps commanders and staffs understand and leverage the pervasive nature of information, its military uses, and its application during all military operations. This function provides JFCs the ability to integrate the generation and preservation of friendly information while leveraging the inherent informational aspects of all military activities to achieve the commander's objectives and attain the end state. [223]

For our research, we will define information operations as *the combination of physical, virtual, and cognitive endeavors undertaken to influence a target audience, organization, or individual to act (or not act) in way that is beneficial to the perpetrator* (combines concepts from [221] and [51]). While we acknowledge that information operations span the physical, virtual, and cognitive domains, we will focus our research on the virtual and cognitive lines of effort. Additionally, while acknowledging that this definition intentionally includes everything ranging from commercial marketing of products to military deception, we will focus on political disinformation that targets another society, or specifically its political structures, leaders, or national security entities. Prier says that this type of information operations "hinges on four factors: (1) a message that fits an existing, even if obscure, narrative; (2) a group of true believers predisposed to the message; (3) a relatively small team of agents or cyber warriors; and (4) a network of automated 'bot' accounts." [195]

For further reading on information operations, see [195] for information operations in social media. See [91] for Russian thought on operations in information space, [198] for Chinese approaches, and [221] for US military doctrine and operational approaches.

1.3 Backdrop: Russian information blitzkrieg

> "Russia is waging the most amazing information warfare blitzkrieg we have ever seen in the history of information warfare."
>
> –Gen Philip Breedlove, NATO Wales summit, September 2014

The Russian connection to modern disinformation operations is highlighted by the fact that the word disinformation didn't enter the English language until the 1980's, and is a direct transliteration of the Russian word дезинформация (pronounced *dezinformatsiya*). Stalin allegedly developed the term, intentionally making it sound Western, and therefore creating disinformation with the very origin of the word [202]. The word was soon used as the title for a division of the KGB that focused on black propaganda. These activities were in line with traditional KGB operations known as 'active measures'. These were described by KGB Major General Oleg Kalugin as "active measures to weaken the West, to drive wedges in the Western community alliances of all sorts, particularly NATO, to sow discord among allies, to weaken the United States in the eyes of the people in Europe, Asia, Africa, Latin America, and thus to prepare ground in case the war really occurs" [3]. This quote highlights one of the critical roles of the Russian information blitzkrieg is to drive wedges in every fissure possible, fracturing a nation or coalition. This includes driving wedges between political parties, between races, between religions, between a nation and its military, and between a nation and its allies. A fractured nation is inherently a less potent nation.

Active measures, which have been discussed in Soviet political circles since the Bolshevik Revolution [167], involves much more than just disinformation. In addition to disinformation, McCauley indicates that active measures included clandestine operations, military deception, provocations, fabrications/forgeries, agents of influence, diversion/sabotage, and "wet affairs" (assassinations, kidnappings, etc). Disinformation involved both overt (propaganda, diplomacy, cultural organizations) and covert (agents of influence, written/oral, front organizations, forgeries) operations [167].

Closely related to "active measures" is reflexive control. Originally developed by Vladimir Lefebvre in 1936 [152], reflexive control is a mathematical model that "uses a specially prepared information message, while hiding the fact that influence is being conducted, in order to control or influence human- or computer-based decision making to voluntarily take a predetermined action." [167]. Reflexive control takes into account the adversary's perception of the situation, goals, decision making process (or algorithm), and the act of making the decision [197]. With details classified by the KGB [167], reflexive control provides the mathematical and psychological theory behind many Russian disinformation operations.

This Russian Propaganda apparatus, long directed at its own society as well as the satellite states of the former Soviet Union, is now being exported abroad. In 2013 General Valery Gersimov identified information warfare as an important aspect of Russian warfare going forward in his now famous article "The Value of Science is in the Foresight" [106]. While the West has viewed this article backwards through the lens of the Ukrainian conflict and has arguably misattributed it as the start of hybrid warfare for Russian armies [22]. His article was his perspective of the Arab Spring as well as US operations in Yugoslavia, Iraq, and Afghanistan. In his view, the Arab Spring and US-led coalitions relied heavily on resources other than conventional military forces to shape these events, including information operations, introducing military forces only at the last minute. Having studied these conflicts, he sought to accelerate ongoing information warfare initiatives, stating "Information warfare opens wide asymmetric possibilities for decreasing the fighting potential of the enemy."

As indicated above, most of the Russian writing on information operations is defensive in nature [11]. Dating all the way back to 1839, European thought indicated that "the political system of Russia could not withstand twenty years of free communication with Western Europe" [78]. This remained true through the Cold War and persists today. In general, they fear the spread of the internet and what they view as biased reporting by Western Media that is eroding their traditional and cultural values [11]. For this reason, Russian leadership view Western information and ideas as a strategic threat and have developed sophisticated ways and means to battle it.

The emerging manifestations of Russian information operations are built on a long history of Soviet era propaganda operations. In 1951 then Yale Law Professor Harold Lasswell summarized the Soviet Propaganda machine by concluding:

> "...the chief strategic aim of [Soviet Propaganda] is to economize the material cost of protecting and extending the power of the Russian elite at home and abroad. Such propaganda is a struggle for the mind of man, from the Soviet point of view, only in the sense that it is a struggle for the control of the material means by which the minds of the masses are believed to be molded. Hence the purpose of Russian propaganda is not peaceful persuasion of the majority of the people in a given country as a

> prelude to taking power. Rather, the task is conceived as that of a minority that must remain an ideological minority until it succeeds in accumulating the material means of obtaining consensus...Soviet propagandists and their agents can lie and distort without inner restraint, for they are largely immunized from the claims of human dignity in any other sense than the dignity of...contributing to the present and future power of the Kremlin elite." [148]

This general approach continues to this day, building a small nuclei while dividing all opposing organizations and institutions, leveraging disinformation at all times. Today, however, technology enables this at a scale and distance unheard of in 1951.

The Russian state is not approaching this haphazardly. Since as early as 2003 the Russian Academy of Sciences has conducted basic research to develop advanced applied mathematical models of information warfare and its application to a society. Their researchers combine social science and mathematical modeling to produce research such as "Mathematical Modeling of Rumors and Information Propagation in Society". While these articles claim to be defensive, their application in offensive operations is assumed.

The Russian military views information warfare broadly and inclusive. In *Conceptual Views Regarding the Activities of the Armed Forces of the Russian Federation in the Information Space*, they define information operations as confronting a state in the information space by damaging information systems, processes, and resources [91]. During the Cold War, some estimate that the Soviet Union had 15000 personnel and up to 5 billion dollars dedicated to active measures [110].

These operations are synchronized by a growing cadre of political technologists. These are leaders, both inside and outside the government, that understand both the human, political, military, and technological domains. Leveraging this 'multi-domain' understanding, they develop and coordinate shaping operations that leverage the cyber and technological domain to affect the social, political, and military domains. As an example, Alexander Malkevich, a Moscow based technologist, recently established the Moscow based www.USAreally.com website in advance of the 2018 mid-term elections in the United States. His mission was to spread twisted narrative and agitation that is picked up by mainstream American news, or at least mainstream news aggregators. The translated personal *description* from his Twitter account states: "Journalist. Media man. A person who is interested in life. And he is not afraid to work in the regions of Russia. And in the name of Russia." This is a political technologist.

According to an article by Kuleshov et al entitled "Information-psychological confrontation in modern conditions: theory and practice" published in the Russian Bulletin of the Academy of Military Sciences, the primary methods of information manipulation from the Russian perspective are:

- "Direct lies for the purpose of disinformation both of the domestic population and foreign societies;
- Concealing critically important information;
- Burying valuable information in a mass of information dross;
- Simplification, confirmation and repetition (inculcation);
- Terminological substitution: use of concepts and terms whose meaning is unclear or has undergone qualitative change, which makes it harder to form a true picture of events;

- Introducing taboos on specific forms of information or categories of news;
- Image recognition: known politicians or celebrities can take part in political actions to order, thus exerting influence on the world view of their followers;
- Providing negative information, which is more readily accepted by the audience than positive." [145]

All of these contribute to increase the Clauswitzian fog of war [11]. Even when not specifically attempting to influence a specific decision, these operations increase the fog of war and make all decisions more difficult. Additionally, these methods are taken with the full understanding that democratic nations must build consensus among citizens. When this fog of disinformation is widely distributed, it is difficult for leaders of democratic nations to build national consensus and act decisively on the world stage. Finally, [11] highlights that information warfare doesn't have clear redlines like conventional or nuclear warfare. Lack of redlines and difficult attribution increases the fog of war. This has complicated and confused our deterrence mindset and policies.

What is the overall strategic goal for Russian disinformation? "...the main aim of information-psychological conflict is regime change in the adversary country (through destroying the organs of government); by means of mass influence on the military-political leadership of the adversary achieving as a minimum an increase in the amount of time available for taking command decisions and lengthening the operational cycle; by means of influence on the mass consciousness of the population – directing people so that the population of the victim country is induced to support the aggressor, acting against its own interests. (Translated)" [145]

Any discussion of Russia must consider the full strategic context (evaluated against all aspects of DIME). While Russia possesses a large nuclear deterrent, a modern and combat experienced military force, a competent and proven information warfare capability, and an increasing diplomatic role in the Middle East and elsewhere, Russia's economy does not make it in the top 10 world economies (as measured by Nominal Gross Domestic Product). Other rising nations (China, Brazil, and India) all outpace Russia economically. In 2019 for example, China's economy was approximately 8 times larger than Russia's economy as measured by nominal gross domestic product.

For further reading on Soviet and Russian active measures, Ajir provides a short summary [11], Gobson testimony to the Senate [110], McCauley provides a thorough examination with description of organization hierarchy and specific examples [167] while Giles provides a summarization [107].

1.4 Notes on China Information Operations

China disinformation is fundamentally different from Russian disinformation. This is partially due to differences in culture. Mattis attempts to summarize the difference between Russian and Chinese information operations with three overarching differences: "set-piece operations vs. playing the man; service-led operations vs. service-facilitated operations; and agents of influence vs. influenced agents" [166]. Russia focuses on set-piece operations focused on strategic ends, whereas China focuses on person-to-person relationships and influencing the individual. Many of China's diplomats and negotiators started their careers in intelligence, which is evidence of

China's focus on personal influence [166]. For Russia, intelligence services are the primary executor of their 'active measures.' For China, intelligence services may at times facilitate, but the leading role goes to more political and diplomatic organizations such as Liaison Department of the PLA's Political Work Department and the United Front Work Department. Finally, Russia depends heavily on intelligence case officers operating as 'agents of influence', whereas China works through 'softer' methods, using 'gatekeepers' to facilitate foreigner engagement in China and thereby influence these individuals [166].

Additionally, Russia operations are often negative, characterized by higher risk and higher reward. China, on the other hand, primarily focuses on flooding social media with a positive narrative about China and the Chinese Communist Party. This is largely performed by the '50-cent Army', government workers and other recruited and trained individuals who are expected to produce positive narrative in social media [198]. In this way, China arguably conducts the largest *astro-turfing* operation in history (*astro-turfing* is when an organization pays for virtual or physical support in order to produce what appears to be an organic grass roots movement).

We have also observed that Russia frequently and adeptly uses internet memes and other means of creative multi-media. China seems to shy away from this, and at times bans memes completely [168]. Russia leverages the evolution of memes to propagate their narrative, whereas China is concerned that the evolution of memes, particularly in their own population, may take a turn and result in negative perception of the CCP. This will be discussed further in Chapter 5.

1.5 Change in the Strategic Center of Gravity

The 20th century dawned with the most symmetric and kinetic wars in the history of warfare, while the 21st century, spring boarding off of decades of Cold War competition, has dawned with numerous asymmetric and non-kinetic conflicts. During World War I, nations sacrificed hundreds of thousands of lives for mere yards of physical terrain. Today, many actors develop complex information operations designed to slowly gain 'yards' in the human domain with ramifications for the physical domain.

Geography still matters today. The United States' two greatest security measures are still called the Pacific and Atlantic Oceans [266]. Crimea was annexed largely because of the strategic importance of its Black Sea Port (as well as energy implications) [41]. Afghanistan instability will persist partly because of its geography [135]. Geography does and always will matter. However, numerous factors, to include technology, have arguably shifted the pendulum toward the human dimension.

This shift toward the human domain was hotly debated inside the US military during the War on Terror. After years of debate, the majority seemed to agree with the quote from a 2009 article in Small Wars Journal: "One of the most profound changes the U.S. military must make to be effective at countering insurgency is to shift strategic centers of gravity from the physical to the human aspects of warfare" [104]. While generally accepted in counter-insurgency environments, it remains to be seen how this shift toward the human domain will change large scale combat operations (LSCO).

This view of the population as the center of gravity took on new meaning in the aftermath of the Arab Spring, as decentralized population movements, enabled by technology, organized and overthrew multiple established autocratic regimes. These actions shocked the world and have been studied by leaders from both the East and the West. These events underscored the power of the human dimension as well as the power of social media to mobilize the masses. Multiple articles in military journals have documented these movements, with a specific focus on the social media that enabled them. Even General Gerasimov's 2013 article in Russia's Military-Industrial Kurier, studied across the West as the genesis of *hybrid* or *grey* warfare, is more a personal reflection of the Arab Spring (as well as the conflicts in Iraq, Afghanistan, and Yugoslavia), than an attempt to create a new type of warfare [22].

Multiple other state and non-state actors observed these changes and began exploring the idea of manipulating these movements through cyberspace. Many of these states and actors already had experience manipulating their own populace or organization through information operations[148], and now sought to extend that experience to other populations and societies. Directly targeting the fabric of society, the true center of gravity of a nation, has massive ramifications for the tactical through strategic levels of war, and is the genesis of this emerging domain of social cyber security.

1.6 Enabling changes

Two changes in human communication and societal information flows have enabled the social cyber threat. First, technology has waived the requirement for physical proximity to influence a society, and the decentralization of information flows has reduced the cost of entry. Fabio Rugge (Italian Institute for International Political Studies) sums this up with the statement "Cyberspace is a powerful multiplier of the destabilizing effects of manipulated information because it allows high connectivity, low latency, low cost of entry, multiple distribution points without intermediaries, and a total disregard for physical distance or national borders. Most importantly, anonymity and the lack of certain attribution of an attack make cyberspace the "domain of ambiguity" [201].

1.6.1 Decentralization

Over the last 30 years we have watched as information flows rapidly decentralized. Historically governments, large organizations, and a few large news outlets controlled most of the formal print, broadcast, and televised news coverage. These organizations controlled the flow of information, and generally distributed it uniformly across a society. With the rise of blogs, microblogs, and social networks, most of the world now consume their information in a non-uniform way on social media [211]. There is now a low cost of entry, financial incentive to create viral content, and anonymity is relatively easy to accomplish. This decentralization has facilitated the entry of external actors with minimal attribution.

Decentralization means that quality control is now decentralized. Fact checking is now conducted at the user level rather than the journalist level. Users, many who grew up in an era where news was largely trusted, are now unprepared to digest news in an era where truth and untruth are mixed, especially if distortions of the truth are designed to validate their own biases.

The decentralization of news has also opened the door to a wide spectrum of news catering to specific audiences and biases. While most large news organizations traditionally centered themselves near the political center, decentralization creates niche news and discussions designed for every part of the political spectrum. Sometimes called "preference bubbles" [250], these bubbles not only create polarization, but also make it relatively easy for information operations to target specific groups within a society.

The traditional journalism business model requires truth. Journalists lose their jobs and news organizations lose business if they are consistently in error. The social media business model, largely focused on overall traffic and advertising, does not rely as much on fact checking (though this is slowly changing, as was observed in the August 2018 stock decline for both Twitter and Facebook, largely attributed to their slow growth while they clean up their platforms).

While recent legislation across the world is trying to find a way to centralize control, in all cases this involves some type of censorship and reduced freedom of speech. In some cases, it could end up in absolute chaos, especially if social media companies are required to provide a platform functionality for people to flag fake or malicious information. If this type of functionality is exposed to users either through the Application Programming Interface (API) or the web/mobile interface, then the same bots that post fake news can now flag all kinds of content as fake at the speed of algorithms, causing exponentially greater damage.

1.6.2 Physical Presence not required

For most of history, influence required physical presence, or at least physical proximity. To influence the conversation of the Roman forum, the heartbeat of Roman society, an actor or proxy must be physically present in the Forum or at least in Rome, clearly identifiable, and active in the conversation. "Cloak and dagger" operations occurred, but even these operations required physical presence. This requirement held true through the first part of the 20th Century, at which time, radio and leaflet operations emerged, not requiring direct physical presence but rather require some level of proximity. Even robust Soviet era propaganda operations were largely restricted to Eastern Europe and Asia due to geographical limitations. The internet has erased this requirement, with most societies interacting in free and open online environments that allow actors to participate from the far corners of the globe with few national borders in the cyber domain.

Those nations that value freedom of speech and open marketplace for opinions and ideas are more vulnerable to these threats [24]. This is most evident by the fact that North Korea, arguably the most closed nation on earth, is still largely immune to social manipulation through the internet. Directly influencing the North Korean society still requires physical presence/proximity.

This vulnerability of open societies is exacerbated by the fact that most of these strategic information efforts are launched on global social media platforms that are privately owned and outside of the direct supervision of governments (though influenced by regulation). While all social media companies censor content on their platform, their motivation is generally focused on improving user experience for the greatest number of people across the world, not national security concerns of any single nation. Choosing sides on any issue is generally bad for business, because is alienates a segment of their customer base. Government censorship of content is assumed to be partisan and violates the freedom of speech espoused by these governments. Third party efforts to censor content have been initiated, but to date these have been narrowly focused and easily circumvented. An example of third party efforts is the *Social Science One* initiative, a creative partnership between academic researchers, private industry, and funding from across the political spectrum that facilitates third party research on social media data while maintaining individual privacy. Efforts like this are still in their infancy.

1.7 Conclusion

> "...a new-generation war will be dominated by information and psychological warfare that will seek to achieve superior control of troops and weapons and to depress opponents' armed forces personnel and population morally and psychologically. In the ongoing revolution in information technologies, information and psychological warfare will largely lay the groundwork for victory." [62]
>
> - Russian *Military Thought*, 2013

Arguably, the greatest strategic weakness for any country is internal, not external. Leaders must understand social cyber security in order to defend these internal weaknesses from external manipulation. We as military leaders must understand that one of the information blitzkrieg lines of effort will be to drive a wedge of distrust between us and the society we defend as well as civil leadership that leads us. An untrusted institution will be under-funded, under-used, and under-performing.

If one of our primary missions is to "sustain American influence abroad" (2018 DoD Mission Statement), then we need to find our role in promoting American values in this international marketplace of beliefs and ideas within a coordinated interagency effort. This influence will range from online interaction to the handshake from a forward deployed platoon leader.

Military leaders must enact policies that enable freedom of maneuver in the relevant information environments. A recent RAND Information Operations report concluded that the Department of Defense must change its policy in order to fully enable ethical maneuver within the information domain [165]. Most social cyber security practitioners (both bot creators and bot defenders) use Application Programming Interface (API's) and open source technology to access and maneuver in this data environment. In other words, API's are the access point for both offensive and defensive social cyber operations. In the military, policies and authorities to access API's are severely restricted for some organizations while not well-defined for others. We need agile policies that enable initiative in a dynamic information environment while protecting the privacy of well-intentioned individuals and remaining within the authorities granted to the Department of Defense.

In summary, we must directly educate our force and indirectly educate our society about the decentralized nature of the modern information environment, the risks that exist, and ways and means to individually vet the facts and opinions that we digest and allow to shape our beliefs and attitudes. We must develop a multi-disciplinary approach to social cyber security. We must build relevant policy that enables social cyber security. We must seek to remove any wedge of distrust artificially driven between our military and the society we defend. We must search for the Department of Defense role in an interagency effort to combat the information blitzkrieg we face today. Social cyber security is a required discipline for the foreseeable future.

The Western World must approach this holistically, and not constrain our aperture to just looking at social media and other cyber-enabled technologies. For example, a huge part of Russian disinformation includes traditional lobbying, purchasing newspapers, and running full page ads in prominent western newspapers [11]. These avenues of approach must be part of our consideration.

Finally, while much of the subsequent chapters focus on identifying and characterizing social cybersecurity threats, we must also initiate research that explores social cybersecurity policy as well as methods to respond to and mitigate the threat. The multi-disciplinary approach to policy must consider the trade-offs of privacy and openness vs. the damage that an undetected threat can have on a society. Policy research must address whether or not “red-lines” are relevant to this threat, and what the response will be if these “red lines” are crossed. These are all recommended lines of investigation for future research.

Chapter 2

Information Warfare *Forms of Maneuver*: BEND

In this chapter we define and describe information warfare "forms of maneuver." These are the information equivalent of the physical forms of maneuver that describe offensive land warfare. The physical forms of maneuver are "...distinct tactical combinations of fire and movement with a unique set of doctrinal characteristics that differ primarily in the relationship between the maneuvering force and the enemy" [122]. The US Army identifies five offensive forms of maneuver: 1) turning movement, 2) envelopment, 3) penetration, 4) infiltration, and 5) frontal assault. In light of these offensive forms of maneuver for land warfare, we developed appropriate forms of maneuver for information warfare. Parts of this Chapter were published in [34] and [35].

Fire and maneuver is the basis for modern ground tactics, and has been used successfully in every major combat engagement since Swedish king Gustavus Adolphus first used it in the 30 Years War (1618 to 1648). At its basic level, fire and maneuver involves one or more tactical units fixing and enemy with direct and indirect fires, while one or more tactical units maneuver to close with and assault the enemy. In the information world the narrative plays a similar role to "fires", while network manipulation and preparation plays the role of "maneuver". Like traditional fire and maneuver, the interplay between network maneuver and targeted narrative must be carefully coordinated. Our overarching premise and largest contribution is defining information maneuver in terms of **BOTH** narrative **AND** networks.

In developing the information warfare forms of maneuver, we build on the *dismiss*, *distort*, *dismay*, and *distract* paradigm introduced by Ben Nimmo at the Atlantic Council's Digital Forensic Research Lab [181]. Nimmo developed 'the 4 D's' to describe emerging Russian information operations. We found these were helpful, but were not sufficient to describe all maneuver we were observing. The 4 D's only describe maneuver in narrative, while ignoring any manipulation of or maneuver in the network. Secondly, Nimmo's four D's only describe negative maneuver (attacking a narrative), while not addressing positive maneuver (supporting a narrative). We therefore developed four B's to describe supporting network maneuver, four E's to describe supporting narrative maneuver, four N's to describe attacking network maneuver, and used Nimmo's four D's to continue to describe attacking narrative maneuver. Combined these become the 'BEND' forms of maneuver.

The BEND forms of maneuver describe how an actor can manipulate the marketplace of beliefs and ideas in order to achieve a desired endstate (this endstate could simply be increased agitation or polarization of the target audience). The BEND Forms of maneuver, with definitions, are provided in Table B.2 (note that each of the four quadrants of the forms of maneuver begin with a letter from the acronym BEND).

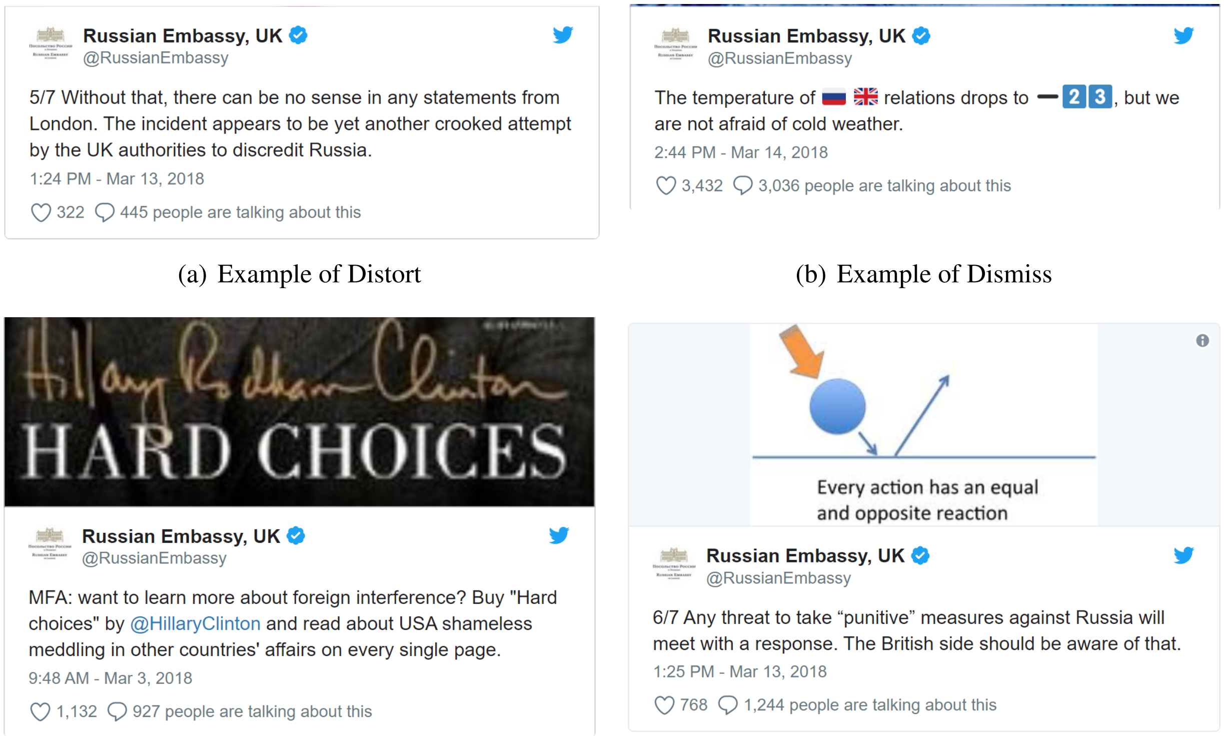
In the following sections we will describe these in more detail and provide examples. We constrained all of our examples of disinformation to examples from overt state actors or, in some cases, alleged proxies of state actors. Note that information operations are not restricted to state actors, and many international as well as domestic actors conduct information operations. These actors include businesses, non-governmental organizations (NGO's), political parties and political groups, as well as terrorist and criminal organizations.

Table 2.1: The BEND model of describing social cyber security *forms of maneuver*

	Network Maneuver		**Narrative Maneuver**	
	Social network manipulation		Knowledge network manipulation	
	Things you can do by effecting “who is talking/listening to whom”		Things you can do by effecting “what is being discussed”	
Constructive	**Back**	actions that increase the importance of the opinion leader	**Engage**	discussion that brings up a related but relevant topic
	Build	actions that create a group or the appearance of a group	**Explain**	discussion that provides details on or elaborates the topic
	Bridge	actions that build a connection between two or more groups	**Excite**	discussion that beings joy, happiness, enthusiasm to the group
	Boost	actions that grow the size of the group or make it appear that it has grown	**Enhance**	discussion that encourages the group to continue with the topic
Destructive	**Neutralize**	actions that limit the effectiveness of opinion leader	**Dismiss**	discussion about why the topic is not important
	Nuke	actions that lead to a group being dismantled	**Distort**	discussion that alters the main message of the topic
	Narrow	actions that lead to the group becoming sequestered from other groups	**Dismay**	discussion about a topic that brings worry, sadness, anger to the group
	Neglect	actions that reduce the size of the group or make it appear the group has grown smaller	**Distract**	discussion about a totally different topic and irrelevant

2.1 Narrative Maneuver

Narrative maneuver is arguably the most common type of information maneuver. In fact, when many people think of information warfare, they naturally envision a battle of narratives where actors compete by creating the most convincing narrative. In information maneuver, actors attack certain narratives while supporting other narratives. They attack narrative through *dismiss*, *distort*, *dismay*, and *distract* maneuvers (these are the original 4 D's developed by Ben Nimmo at the Digital Forensic Labs). These have been extremely helpful in characterizing Russian information operations. Examples of these forms of maneuver are provided in Figure 2.1.



Russian Embassy, UK
@RussianEmbassy

5/7 Without that, there can be no sense in any statements from London. The incident appears to be yet another crooked attempt by the UK authorities to discredit Russia.

1:24 PM - Mar 13, 2018

♡ 322 💬 445 people are talking about this

(a) Example of Distort

Russian Embassy, UK
@RussianEmbassy

The temperature of 🇷🇺 🇬🇧 relations drops to ➖2️⃣3️⃣, but we are not afraid of cold weather.

2:44 PM - Mar 14, 2018

♡ 3,432 💬 3,036 people are talking about this

(b) Example of Dismiss

Russian Embassy, UK
@RussianEmbassy

MFA: want to learn more about foreign interference? Buy "Hard choices" by @HillaryClinton and read about USA shameless meddling in other countries' affairs on every single page.

9:48 AM - Mar 3, 2018

♡ 1,132 💬 927 people are talking about this

(c) Example of Distract

Russian Embassy, UK
@RussianEmbassy

6/7 Any threat to take "punitive" measures against Russia will meet with a response. The British side should be aware of that.

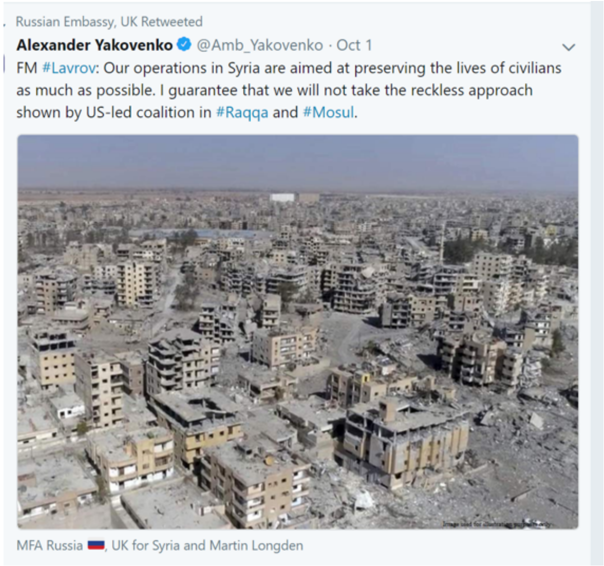
1:25 PM - Mar 13, 2018

♡ 768 💬 1,244 people are talking about this

(d) Example of Dismay

Figure 2.1: Examples of negative narrative maneuver deployed by the Russians in the aftermath of the poisoning of Sergei and Yulia Skripal in Britain

We felt these are not sufficient since not all information efforts attack certain beliefs, ideas, and narratives. In other words, not all information operations are destructive. Many efforts are trying to support or back other beliefs, ideas, and narratives. In this way, they are constructive. This is particularly true in the case of China information operations, which traditionally executes a massive astro-turfing operation (known as the '50 cent Army') with the sole purpose of constructing and supporting pro-China and pro-CCP narratives. We therefore added the constructive narrative maneuvers of *engage*, *explain*, *excite*, and *enhance*. Note that we call these constructive and supporting in relation to the narrative they're supporting, even if that narrative is considered negative, false or destabilising. Examples of these constructive narrative maneuvers are provided in Figure 2.2.



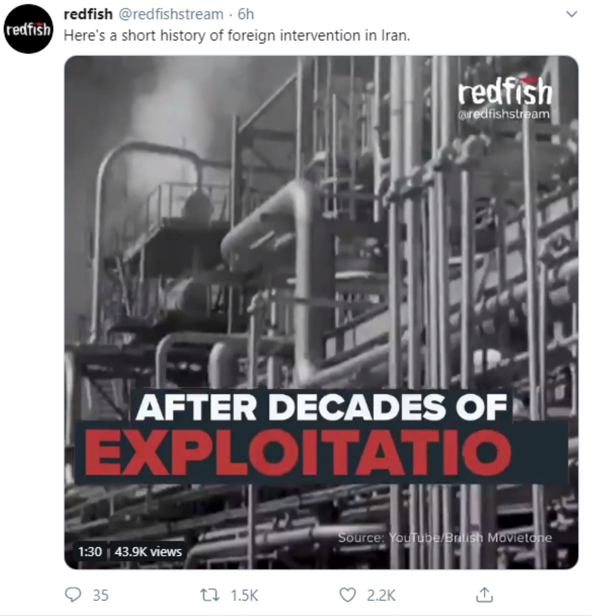
Russian Embassy, UK Retweeted

Alexander Yakovenko @Amb_Yakovenko · Oct 1

FM #Lavrov: Our operations in Syria are aimed at preserving the lives of civilians as much as possible. I guarantee that we will not take the reckless approach shown by US-led coalition in #Raqqa and #Mosul.

MFA Russia, UK for Syria and Martin Longden

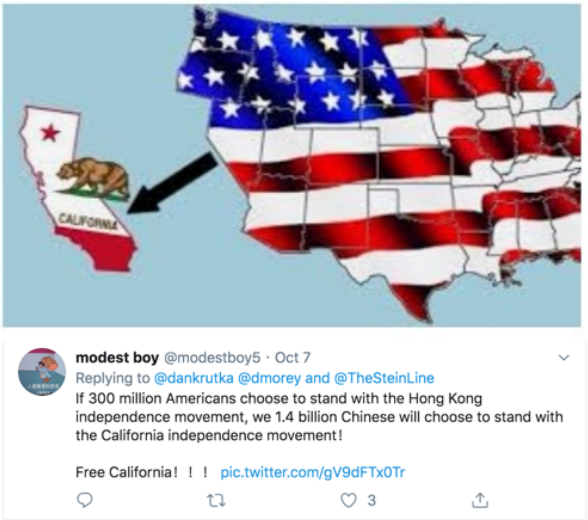
(a) Example of Engage: Russian overt propaganda tries to engage the Middle East audience and offer a biased comparison of US and Russian approaches to ISIS



redfish @redfishstream · 6h

Here's a short history of foreign intervention in Iran.

(b) Example of Explain: Redfish, a Berlin-based media company backed by the Kremlin, creates a short documentary attempting to highlight Western exploitation of Iran



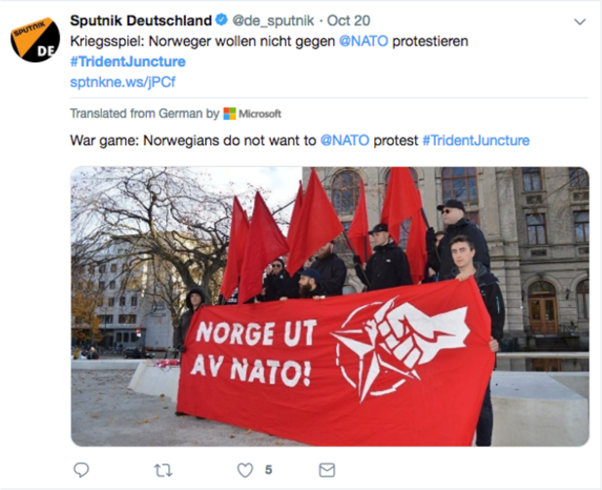
modest boy @modestboy5 · Oct 7

Replying to @dankrutka @dmorey and @TheSteinLine

If 300 million Americans choose to stand with the Hong Kong independence movement, we 1.4 billion Chinese will choose to stand with the California independence movement!

Free California! ! ! pic.twitter.com/gV9dFTx0Tr

(c) Example of Excite: Chinese proxy attempts to anger and excite American support for Hong Kong protests by comparing it to support for a California Independence Movement



Sputnik Deutschland @de_sputnik · Oct 20

Kriegsspiel: Norweger wollen nicht gegen @NATO protestieren #TridentJuncture
sptnkne.ws/jPCf

Translated from German by Microsoft

War game: Norwegians do not want to @NATO protest #TridentJuncture

(d) Example of Enhance: Russian State Sponsored Media in Germany attempts to enhance and amplify grassroot Scandinavian protests against NATO

Figure 2.2: Examples of constructive state sponsored information maneuver that supports specific narratives

These forms of maneuver can be deployed with various computational approaches in modern social media environments. Examples of computational approaches that leverage narrative maneuver include:

1. Misdirection: Introduction of unrelated divisive topics into a thread in order to shift the conversation
2. Hashtag-latching: Tying content and narratives to unrelated trending topics and hashtags
3. Smoke screening: Spreading content (both semantically and/or geographically) that masks other operations
4. Thread-jacking: Aggressively disrupting or coopting a productive online conversation

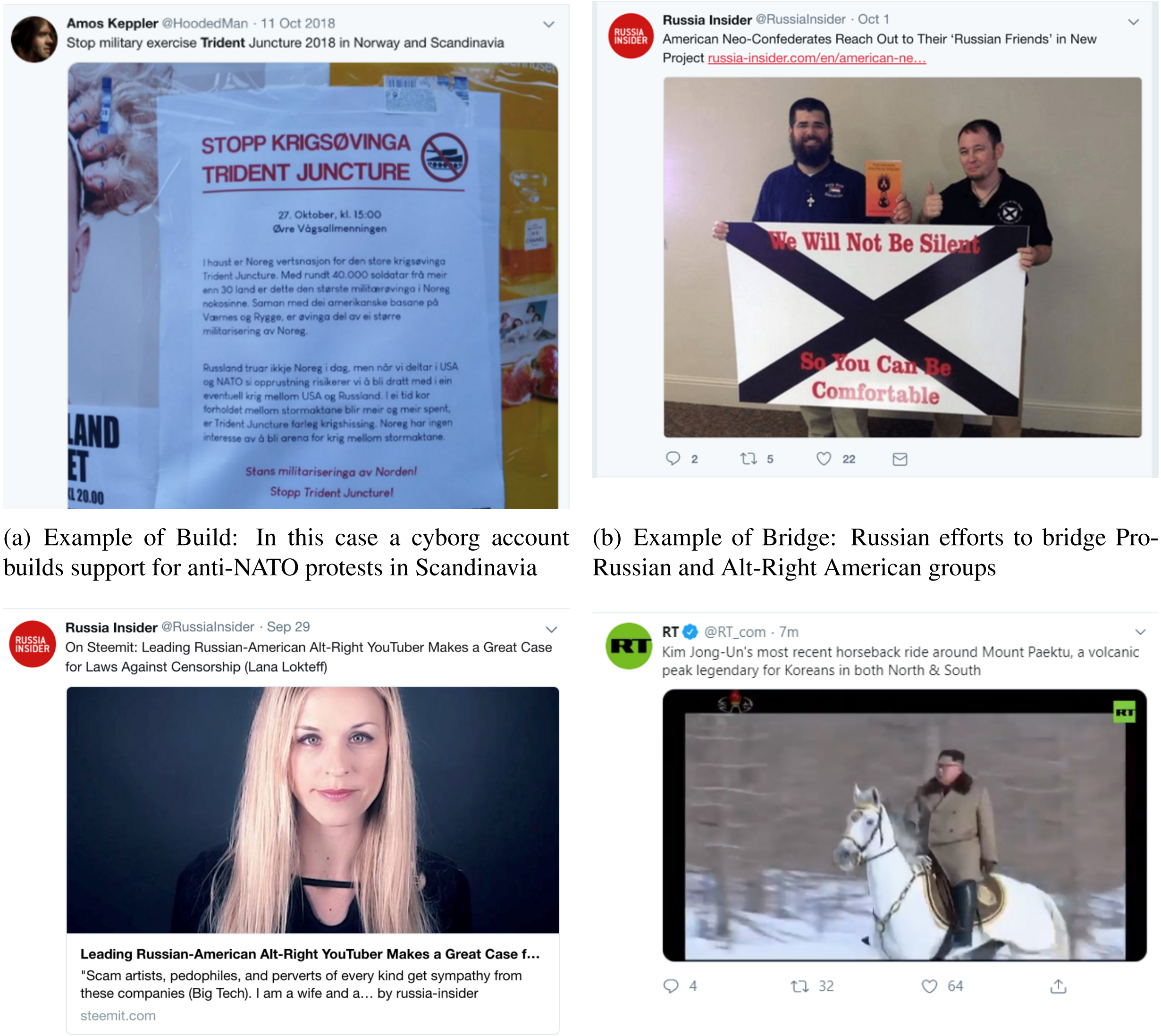
2.2 Network Maneuver

In the social cyber domain, an adversary can manipulate the network as well as the narrative. These networks can be social networks (Sarah is *friends* with Peter), conversation networks (Sarah *replies* to Peter), or informational networks (Sarah and Peter both share the hashtag #NATO). This maneuver includes building or attacking links: links between individuals, links between groups, or links between topics. It can involve building, bridging or attacking groups, both social groups or topic groups. It can involve identifying and supporting or attacking influencers, where influencers are celebrities, politicians, bloggers, bots, cyborgs, or captivating multimedia that are influential in the marketplace of beliefs and ideas. All this manipulation of the network is often necessary to prepare it for the narrative.

Network maneuver is the manipulation of the actual network. In these maneuvers an adversary maps a social network (once again realizing that an online social network is the projection of social and conversational links in the cyber dimension). The constructive network maneuvers are back, build, bridge, and boost. These are defined in Table B.2 and examples are provided in Figure 2.3. The destructive network maneuvers are also defined in Table B.2 and examples are provided in Figure 2.4.

These network forms of maneuver can be deployed with various computational and psychological approaches in the modern social media environments. Examples of computational approaches that support network maneuver include:

1. Opinion Leader Co-opting: gaining access and acknowledgement from an online opinion leader and leveraging their influence to spread narrative.
2. Community Building: Build a community around a topic, idea, or hobby and then inject narrative into this group. This was accomplished in Ukraine by building communities of young men around adult content sharing accounts, and then inject anti-Ukrainian and pro-Russian rhetoric into these networks.
3. Community bridging: In this case the adversary will identify two communities, A and B. They would like to inject ideas of group B into group A. They do this by first infiltrating group A, then slowly adding retweets or sharing ideas from group B, bringing the ideas of group B into group A.
4. False generalized other: This technique relies on the false notion that people have that a



Amos Keppler @HoodedMan · 11 Oct 2018
Stop military exercise **Trident** Juncture 2018 in Norway and Scandinavia

STOPP KRIGSØVINGA TRIDENT JUNCTURE

27. Oktober, kl. 15:00
Øvre Vågsallmenningen

I haust er Noreg vertsnasjon for den store krigsøvinga Trident Juncture. Med rundt 40.000 soldatar frå meir enn 30 land er dette den største militærøvinga i Noreg nokosinne. Saman med dei amerikanske basane på Værnes og Rygge, er øvinga del av ei større militarisering av Noreg.

Russland truar ikkje Noreg i dag, men når vi deltar i USA og NATO si opprustning risikerer vi å bli dratt med i ein eventuell krig mellom USA og Russland. I ei tid kor forholdet mellom stormaktane blir meir og meir spent, er Trident Juncture farleg krigshissing. Noreg har ingen interesse av å bli arena for krig mellom stormaktane.

Stans militariseringa av Norden!
Stopp Trident Juncture!

(a) Example of Build: In this case a cyborg account builds support for anti-NATO protests in Scandinavia

Russia Insider @RussiaInsider · Oct 1
American Neo-Confederates Reach Out to Their 'Russian Friends' in New Project russia-insider.com/en/american-ne…

(b) Example of Bridge: Russian efforts to bridge Pro-Russian and Alt-Right American groups

Russia Insider @RussiaInsider · Sep 29
On Steemit: Leading Russian-American Alt-Right YouTuber Makes a Great Case for Laws Against Censorship (Lana Lokteff)

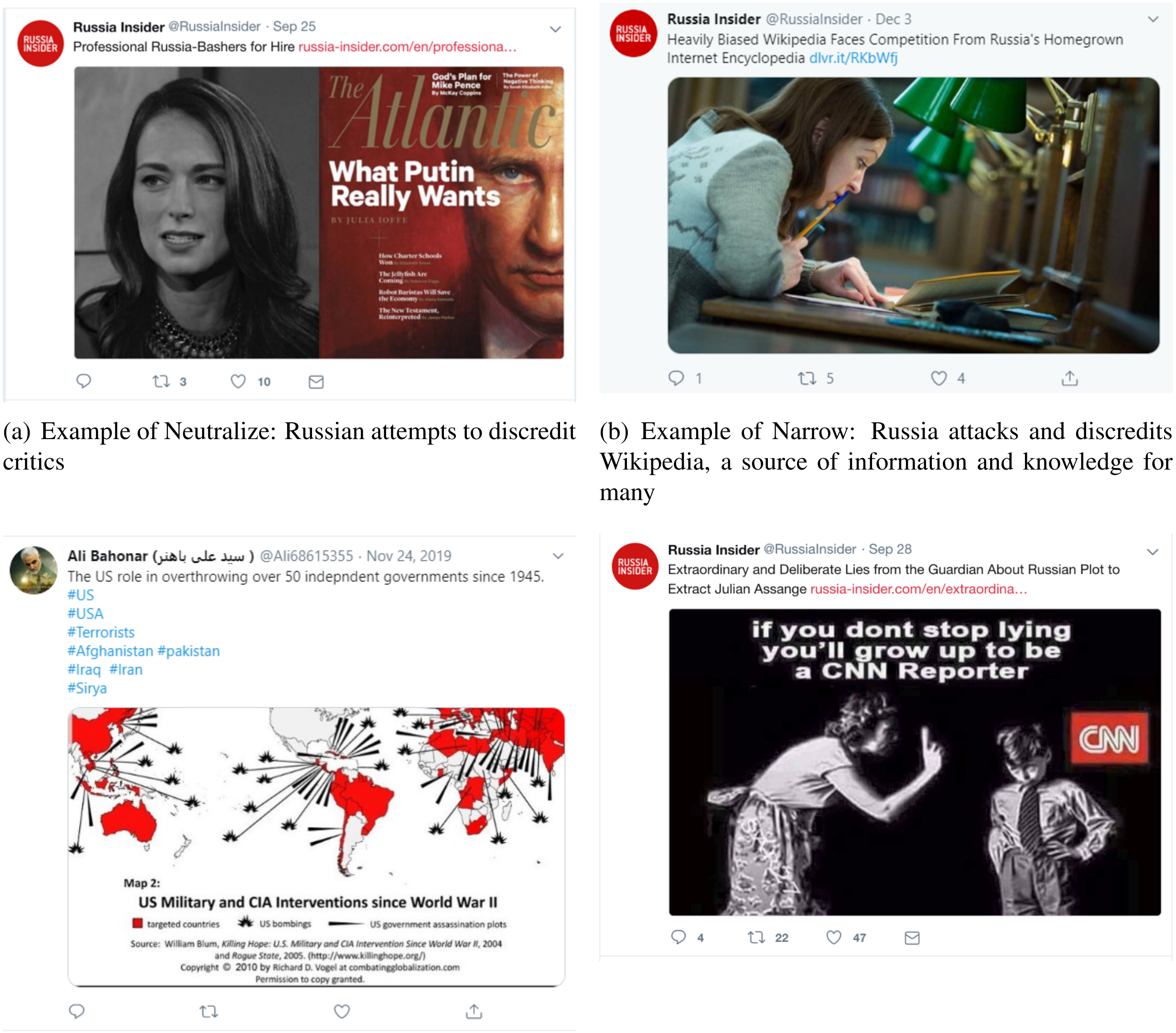
Leading Russian-American Alt-Right YouTuber Makes a Great Case f…
"Scam artists, pedophiles, and perverts of every kind get sympathy from these companies (Big Tech). I am a wife and a… by russia-insider
steemit.com

(c) Example of Boost: Russian efforts to boost Alt-Right and Anti-Censorship campaigns (Censorship restricts Russian freedom of maneuver in information space)

RT @RT_com · 7m
Kim Jong-Un's most recent horseback ride around Mount Paektu, a volcanic peak legendary for Koreans in both North & South

(d) Example of Back: Russian state sponsored media back other authoritarian regimes

Figure 2.3: Examples of Network Maneuver that supports network connections and groups



Russia Insider @RussiaInsider · Sep 25
Professional Russia-Bashers for Hire russia-insider.com/en/professiona...

Russia Insider @RussiaInsider · Dec 3
Heavily Biased Wikipedia Faces Competition From Russia's Homegrown Internet Encyclopedia dlvr.it/RKbWfj

(a) Example of Neutralize: Russian attempts to discredit critics

(b) Example of Narrow: Russia attacks and discredits Wikipedia, a source of information and knowledge for many

Ali Bahonar (سید علی باهنر) @Ali68615355 · Nov 24, 2019
The US role in overthrowing over 50 indepndent governments since 1945.
#US
#USA
#Terrorists
#Afghanistan #pakistan
#Iraq #Iran
#Sirya

Map 2:
US Military and CIA Interventions since World War II
targeted countries
US bombings
US government assassination plots
Source: William Blum, Killing Hope: U.S. Military and CIA Intervention Since World War II, 2004 and Rogue State, 2005. (http://www.killinghope.org/)
Copyright © 2010 by Richard D. Vogel at combatingglobalization.com
Permission to copy granted.

Russia Insider @RussiaInsider · Sep 28
Extraordinary and Deliberate Lies from the Guardian About Russian Plot to Extract Julian Assange russia-insider.com/en/extraordina...

(c) Example of Neglect: Iranian proxy account attempts to Reduce Number of US Partners and Allies

(d) Example of Nuke: Russian effort to discredit mainstream media in the United States in order to open the market for alternative news

Figure 2.4: Example of Network Maneuver that attacks certain network connections and groups

given idea represents the consensus of the masses, and therefore should be their idea and belief.

Because these network maneuvers aren't as intuitive as the narrative maneuver, we will provide several specific examples of these maneuvers being deployed around world events. Benigni et al observed an alleged Russian campaign to *build* networks of young men around adult content on social media, and then introduce pro-Russian narrative and call to arms once the group was built [26]. Several information efforts have been observed that appear to bridge alt-right political groups with mainstream protestant and evangelical online communities [32].

2.3 Bots As Force Multipliers

Within the context of information operations, bots are increasingly used as force multipliers. They leverage machine learning and artificial intelligence to conduct targeted and timely information transactions at scale, while leaving critical nuanced dialogue to human operators (in this context sometime referred to as trolls). We will discuss bots in detail in Chapter 3 and 4, but want to discuss them here in regards to information maneuver.

Bots are used for a wide variety of reasons, creating effects that can are positive, nuisance, or malicious. Some examples of positive bots include personal assistants and accounts that notify the public of natural disasters. Nuisance bots distribute spam with content ranging from commercial advertising to adult content. Malicious bots are typically involved in propaganda [161], suppression of dissent [240], intimidation [159], and network infiltration/manipulation [26].

A bot is defined as social media account that uses a computer to automate social media tasks. For example, in the Twitter environment, a bot account can automatically *tweet*, *retweet*, *follow*, *friend*, *reply*, *quote*, and *like*. The bot creator can use creative means to generate content, either 'scraping' (and automatically summarizing) from elsewhere on the web, retweeting existing content, manipulating existing content from other human users, or creating their own content through a combination of human input and artificial intelligence. Having created content, the bot creator can manipulate tweet timing to appear human (or if appearing human is not critical to the operation, can conduct 1000's of actions around the clock). Finally, these bots are often deployed in bot nets (sometimes called bot 'armies' or 'coordinating' bots) where they *friend*, *follow*, and otherwise promote each other to appear popular.

Although we often attempt to classify an account as *bot* or *human*, there is often a spectrum of automated involvement with an account. Many accounts are not strictly automated (all transactions executed by a computer). These accounts have human intervention to contribute nuanced dialogue while a computer executes tasks at scale in the background.

When combined with artificial intelligence, these bots conduct sophisticated operations at scale at the speed of algorithms.

2.4 Concluding Comments

In addition to dividing these by narrative or network maneuvers, we also separate them into constructive or destructive maneuvers. Constructive maneuvers build or strengthen narratives

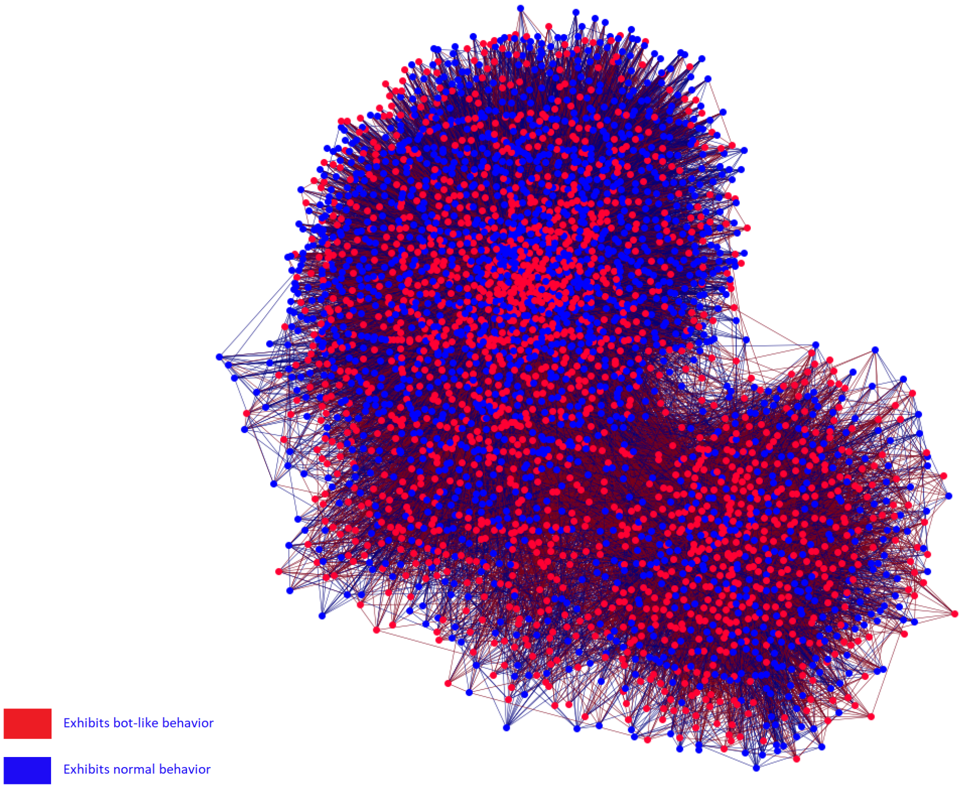


Figure 2.5: Bot involvement in the core Twitter political conversation surrounding the 2018 National Elections in Sweden.

and networks, and destructive maneuvers attack narratives and networks. Note that just because there's a constructive maneuver doesn't mean that the narrative or network is good or positive. For example, building a neo-nazi network or enhancing a racist narrative isn't positive or good, but these maneuvers are defined as 'constructive' since they're designed to grow the respective narrative and network.

Identifying BEND forms of maneuver is an important step in analyzing the main engine of an information operations campaign that is linking content to a manipulated target audience. The BEND forms of maneuver are therefore important analysis tools used in the Sketch-IO framework discussed in Chapter 7. Early metrics that help identify BEND forms of maneuver are being implemented in ORA-PRO, a joint software venture between Carnegie Mellon University and the Netanomics company.

Chapter 3

Bot Detection

Automated social media *bots* have existed almost as long as the social media platforms they inhabit. Although efforts have long existed to detect and characterize these autonomous agents, these efforts have redoubled in the recent months following sophisticated deployment of bots by state and non-state actors. This research will study the differences between human and bot social communication networks by conducting an account snowball data collection, and then evaluate network, content, temporal, and user features derived from this communication network in several bot detection machine learning models. We will compare this model to the other models of the *bot-hunter* toolbox as well as current state of the art models. In the evaluation, we will also explore and evaluate relevant training data. Finally, we will demonstrate the application of the *bot-hunter* suite of tools in Twitter data collected around the Swedish National elections in 2018.

3.1 Introduction

Automated and semi-automated social media accounts have been thrust into the forefront of daily news as they became associated with several publicized national and international events. These automated accounts, often simply called *bots* (though at times called *sybils*), have become agents within the increasingly global marketplace of beliefs and ideas. While their communication is often less sophisticated and nuanced than human dialogue, their advantage is the ability to conduct timely informational transactions effortlessly at the speed of algorithms. This advantage has led to a variety of creative automated agents deployed for beneficial as well as harmful effects. While their purpose, characteristics, and "puppet masters" vary widely, they are undeniably present and active. Their effect, while difficult if not impossible to measure, is tangible.

Automated and semi-automated accounts are used for a wide variety of reasons, creating effects that can be positive, nuisance, or malicious. Examples of positive bots include personal assistants and natural disaster notifications. Nuisance bots are typically involved in some type of 'spam' distribution or propagation. The spam content ranges from commercial advertising to the distribution of adult content. Malicious bots are involved in propaganda [161], suppression of dissent [240], and network infiltration/manipulation [26].

Malicious bots have recently gained wide-spread notoriety due to their use in several major international events, including the British Referendum known as “Brexit” [130], the American 2016 Presidential Elections [39], the aftermath of the 2017 Charlottesville protests [109], the German Presidential Elections [180], the conflict in Yemen [159], and recently in the Malaysian presidential elections [16]. These accounts attempt to propagate political and ideological messaging, and at times accomplish this through devious cyber maneuver.

As these bots are used as one line of effort in a larger operation to manipulate the marketplace of information, beliefs, and ideas, their detection and neutralization become one facet of social cyber security. This chapter and the next chapter will discuss social media bots and bot detection in the context of social cybersecurity. Parts of this chapter have been published in [30, 31, 33, 37].

We are also seeing an increasing number of accounts that we call “bot assisted” or “hybrid” accounts (also at times called “cyborg” accounts). Although researchers often attempt a binary classification of *bot* or *human*, the reality is that there is a spectrum of automated involvement with an account. Many accounts are no longer strictly automated (all content and social transactions executed by a computer). These accounts will have human intervention to contribute nuanced messaging to two-way dialogue, but will have a computer executing a variety of tasks in the background. Grimme et al. [114] discusses this spectrum in detail, describing how ‘social bots’ are created, used, and how ‘hybridization’ can be used to bypass detection algorithms (in their case successfully bypassing the ‘Botornot’ algorithm discussed later in this chapter).

We hypothesize that bots are not involved in social networks and social communication in the same way that humans are, and that this difference is measurable. Like other complex systems (natural ecosystems, weather systems, etc), social interaction and relationships are the result of myriads of events and stimuli in both the real and virtual worlds. Because bots lack real world engagement and social environments, they embed in different networks than humans.

Many bots are programmed to interact with each other as a bot network, and attempt to interact with humans, but many features of these interactions will be ‘robotic’. Even ‘hybrid’ accounts will have some level of artificial and inorganic structure and substance in their communications. This area of bot detection in Twitter is largely unexplored, primarily because the rich network data (both the friends/followers network as well as their conversational network) are very time consuming to collect. We therefore set out to collect the data to characterize the social network(s) and social conversation(s) that a Twitter account participates in, describe these networks with various network metrics, leverage these rich network metrics in traditional machine learning models, and evaluate whether the time involved creates substantial value.

3.1.1 Research questions

1. Do bot Twitter accounts have fundamentally different conversational network structures than human managed accounts?
2. Do the conversations that surround bot accounts diverge from human conversations in general substance and timing?
3. Can the measured differences between bot and human conversation networks lead to increased accuracy in bot detection?

This chapter will begin by discussing past bot detection techniques, as well as summarize historical techniques for extracting features from network structures. Next, we discuss our data collection, data annotation, and methodology for creating ego-network metrics. We describe training and testing our bot-hunter machine learning algorithms and present our results. We construct an evaluation to compare all bot-hunter models against the state of the art. Finally, we will demonstrate the application of the bot-hunter suite of tools in the 2018 Swedish National elections, providing a possible workflow to open source intelligence practitioners.

This chapter is an extension of [33], with a focus of extending the feature space beyond network metrics to include content and temporal metrics of the larger ego network. Several of these features are novel, including a cascaded classifier that identifies the portion of alters that are likely bots, the portion of alters that don't have normal daily rhythms, as well as the portion of ego network that produces tweets that are more popular than the account itself. All of these have been documented as attributes of bots, and we've coded them into features in this algorithm. Additionally, we used the larger models to explore several new bot data sets. Finally, this extension will compare all of the bot-hunter suite of tools against state of the art models.

3.2 Related Work

3.2.1 Understanding Data Tiers

In earlier research our team proposed a tiered approach to bot detection [30] that mirrors the data tiers introduced below. This tiered approach creates a flexible bot-detection "tool-box" with models designed for several scenarios and data granularities. Tier 0 builds models on a single entity (usually a tweet text or *user* screen name). Tier 1 builds models based on features extracted from the basic Tweet object (and associated *user* object). Tier 2 extracts features from a users' timeline, and Tier 3 (explained in this paper) builds features from the conversation surrounding a user. Higher tier models are generally more accurate but consume more data and are therefore computationally expensive. Some research requires bot detection at such a scale, that models based on Tier 0 or Tier 1 are the only feasible option. At other times, highly accurate classification of a few accounts is required. In these cases, models based on Tier 2 or Tier 3 data are preferred. This paper proposes an approach to Tier 2-3 bot detection that builds on the previous Tier 0 [31] and Tier 1 [30] research and relies heavily on network metrics collected through single seed snowball sampling. We will view past research in bot detection through the lens of these tiers.

Since the early efforts to conduct bot/spam detection, numerous teams have developed a variety of models to detect these. While similar, these models will differ based on the underlying data they were built on (for example many community detection and clickstream models were developed for Facebook, while the overwhelming majority of models built on Twitter data use Supervised and Unsupervised Machine learning [6]). Even in Twitter bot detection, these models can be grouped by either the models/methods or by the data that they use. We have provided Table 3.2 to outline the connection between past models and the data that they use.

Adewole et al. [6] reviewed 65 bot detection articles (articles from 2006 - 2016) and found that 68% involved machine learning, 28% involved graph techniques (note that these include some machine learning algorithms that rely heavily on network metrics), and 4% involved crowd-

Table 3.1: Four *tiers* of Twitter data collection to support account classification (originally presented in [30])

Tier	Description	Focus	Collection Time per 250 Accounts	# of Data Entities (i.e. tweets)
Tier 0	Tweet text only	Semantics	N/A**	1
Tier 1	Account + 1 Tweet	Account Metadata	$\sim$ 1.9 sec	2
Tier 2	Account + Timeline	Temporal patterns	$\sim$ 3.7 min	200+
Tier 3	Account + Timeline + Friends Timeline	Network patterns	$\sim$ 20 hrs	50,000+

** This tier of data collection was presented by [144] and assumes the status text is acquired outside of the Twitter API

Table 3.2: Table of Twitter Bot Detection Models and the Data that They Use

		Machine Learning		
Data	Community Detection	Supervised	Unsupervised	Crowd Sourcing
Tier 0 Text		[31, 144]	[151]	[247]
Tier 1 + Profile		[64, 149]	[103]	
Tier 2 + History		[239]	[60]	
Tier 3 + Snowball	[26]	No Known Research		
Stream		[14, 40]		

sourcing. Below we will summarize the salient works under each of these modeling techniques.

3.2.2 Machine Learning Techniques

As noted above, Twitter bot detection has primarily used Machine Learning models. The *supervised* machine learning models used for bot detection include Naïve Bayes [64], Meta-based [149], SVM [151], and Neural Network [144]. The *unsupervised* machine learning models used include hierarchical [151], partitional [103], PCA-based [242], Stream-based [172], and correlated pairwise similarity [60]. Most of these efforts leverage data collected from the basic tweet object or user object (what we would define as a *Tier 0* or *Tier 1* model).

In 2014, Indiana University launched one of the more prominent supervised machine learning efforts with the *Bot or Not* online API service [74] (the service was recently rebranded to *Botometer*). This API uses 1,150 features with a random forest model trained on a collage of labeled data sets to evaluate whether or not an account is a bot. *Botometer* leverages network, user, friend, temporal, content, and sentiment features with Random Forest classification [93].

In 2015 the Defense Advanced Research Projects Agency (DARPA) sponsored a Twitter bot detection competition that was titled “The Twitter Bot Challenge” [226]. This four week competition pitted four teams against each other as they sought to identify automated accounts that had infiltrated the informal Anti-Vaccine network on Twitter. Most teams in the competition tried to use previously collected data (mostly collected and tagged with *honey pots*) to train detection algorithms, and then leverage tweet semantics (sentiment, topic analysis, punctuation analysis, URL analysis), temporal features, profile features, and some network features to create a feature space for classification. All teams used various techniques to identify initial bots, and then used traditional classification models (SVM and others) to find the rest of the bots in the data set.

3.2.3 Other Techniques

Several other novel bot detection methods exist outside of machine learning and network based approaches. Wang et al. [247] investigated the idea of Crowd Sourcing bot detection. While showing limited success, it was costly at scale, and usually required multiple workers to examine the same account. Another unique type of unsupervised learning involves algorithms that find and label correlated accounts. Most bots are not deployed by themselves. Even if not deployed as a united bot-net, many *bot herders* often task multiple bots to perform the same operations. Chavoshi, Hamooni, and Mueen [60] leveraged the semantic and temporal similarity of accounts to identify bots in an unsupervised fashion, creating the *Debot* model which we will compare against in our results section.

3.2.4 Network based techniques

Networks are an extremely important part of bots, bot behavior and bot detection. Aiello et al. [10] discusses the impact of bots on influence, popularity, and network dynamics. Adewole et al. [6] highlights that network features are robust to criminal manipulation.

One approach to leveraging network structure involves community-based bot/sybil detection. While community detection has been effectively implemented on Facebook [262] and Seino Weibo [157], it has only recently been used on Twitter Data due the strict friend/follower rate limiting discussed above. Only recently has Benigni et al. [28] used dense subgraph detection to find extremists and their supporting bots in Twitter.

Most research that uses networks for bot detection with Twitter Data are in fact creating network based metrics and introducing these features in traditional machine learning models. As discussed below, the most challenging part of this type of research is focused on how to build networks from limited data. The closest works to ours were performed by [40] in 2013 and [14] in 2016. Both research efforts used network features along with profile and temporal features from a Twitter Sample Stream without any snowball sampling enrichment. They created an egocentric network that involved ego, alters, with links between alters for both following and mention ego centric networks. Having done this, they calculated content, profile, and social interaction features. Their network features were restricted to centrality measures, density measures, and weak and strongly connected components. A similar earlier work by [40] attempts to use community features (number of communities, core/periphery, foreign in/out degree, etc). This was applied to both Facebook data and the Enron email data (not to Twitter).

Additionally, the *Botometer* algorithm leverages some network features extracted from the user timeline. This includes metrics on the *retweet* network, *mention* network, and *hashtag* co-occurrence network. The metrics include density, degree distributions, clustering coefficient, and basic network characteristics. The *Botometer* algorithm does not conduct a snowball collection of *friends* or *followers*, but does appear to collect user objects for accounts found in the timeline as a *retweet* or *mention* [93].

3.2.5 Building Networks with Twitter Data

As noted above, however, it is difficult to quickly build comprehensive network structure with Twitter data due the Twitter API rate limits, primarily associated with collecting friend/follower ties. Researchers have generally used one of two methods to build limited networks.

The first method is used if the research team has a large sample or stream. These samples may be random (collected from the 1% Twitter Sample) or they may be associated with an event or theme (i.e. collecting all Tweets that have a given hashtag like #hurricanesandy). These researchers then build ego-centric networks from this stream, without collecting any additional data from the Twitter API. This has the advantage of speed, and doesn't suffer from issues getting data for suspended accounts. This method, however, will only model a small portion of an account's activity and network. A 1% sample will arguably contain marginal activity for a given account, and even topical streams will only contain a small part of an account's activity, given that they are involved in multiple topics and discussions. These small samples may not be rich enough to serve as strong features for machine learning.

The second method that researchers use is to only collect the users *timeline* (history of tweets, up to the last 3200). They then build an ego-centric network from this data (variously using replies, retweets, hashtags, urls, and mentions to build networks). This is much richer than the first method, contain all of the user's activity, but still lacks any information beyond that individual, providing the limited star graph illustrated in Figure 3.1. It doesn't contain the larger conversation(s) that they are participating in. Additionally, a bots's *timeline* is completely managed by the bot *puppet master*, and therefore can be manipulated to avoid detection.

To date our team has not found supervised learning bot detection research that leverages extensive snowball sampling to build ego networks.

3.2.6 Extracting Features from Social Networks

Evaluating network centrality measures, started by Bavelas in 1948 [25] and effectively clarified by Freeman in 1978 [100], has long been an important metric for evaluating both nodes and networks. According to Freeman, network-level centrality metrics measure the "compactness" of the network. Our model includes several network centrality measures: degree centrality, k-betweenness [21], and eigenvector centrality [136] are used to measure differing "compactness" between human and bot conversation networks.

Several seminal works describe the importance of triadic relationships in social networks [57, 123] and as a foundation for measuring network clustering and groups [129]. The fact that the study of triadic relationship has almost exclusively been contained within the study of social interaction provides evidence that these observed triadic relationships are unique to human behavior. We have therefore included several features based on these triadic relationships, including the full triadic census [128], number of Simmelian Ties [81, 142], and clustering coefficient. We also included reciprocity based on [174]'s examination of reciprocity in online social networks.

In addition to finding network centrality and triadic structures, network community detection has been an important aspect of network characterization, and is still an active research area. Current group detection techniques generally fall into traditional methods, divisive methods, modularity based methods, statistical inference methods, and dynamic methods [97]. Our community detection features leverage Louvain Clustering [46], which is based on modularity optimization.

Our approach uses network sampling in order to restrict the time of computation. While research in network sampling started in the 1970's with work from [98] and others, the emergence of Online Social Networks (OSN's) increased the size of networks and the need for sampling. Our approach to sampling ego networks was informed by [108]. Our sampling uses breadth-first-search (BFS) on the target node. The known bias of BFS is eliminated because we are only conducting 2 hops from the target (only includes friends of alters).

Finally, the study of ego networks is a special branch of social network analysis that is relevant to our study. In 1972, [113] presented the classic concept of the "Strength of Weak Ties" in ego networks, which [52] clarified is more due to the structural location of ties, and can be measured by effective size, efficiency, and constraint. This informed our use of ego network effective size in our features space. Additionally, centrality of ego-networks was explored by [101] in 1982 informing our use of betweenness in the feature space.

3.2.7 Contributions of this work

While we discuss above several other research attempts to use network metrics in a bot detection feature space, these have largely relied on the mention network extracted from any Twitter query/stream. Ego-centric networks built on a single stream/query arguably contain only a small subset of the overall account ego network. Researchers have not attempted to build this ego network based on snowball sampling [111] with a seed node since this requires significant time given the extent of the data and the strict API rate limits that Twitter imposes on friend/follower data. Our research has taken the time to build this rich conversational network in a novel way, and then evaluate whether the time and effort render sufficient value.

Having built this extensive network for every account in question, this work attempts to fully exploit all available features, going above and beyond just structural features. These additional features include content, temporal, and user summary features. Adding the full range of additional features allows us to fully evaluate the increased accuracy against the additional computational cost.

This work additionally creates and explores bot detection metrics that require greater effort and sophistication to circumvent. Currently, bot-herders can circumvent current algorithms by changing their screen name, adding account meta-data, spending additional time selecting a unique profile picture, and creating a more realistic tweet inter-arrival time. They can also deploy bots in bot networks, therefore artificially manipulating friend/follower values to appear like they are popular. However, it will arguably require significantly more sophistication to change the centrality, components, or triadic relationships in the conversations that they participate in. By increasing the cost to deploy and operate bots, it may economically force "bot-herders" out of their devious market.

Finally, the *bot-hunter* framework builds on the multi-tiered bot detection approach that we introduced in [30]. This multi-tiered approach provides researchers and government or non-governmental agencies with a "tool-box" of models designed for different classes of bots as well as different scales of data (designed for either high volume of high accuracy). This multi-tiered approach acknowledges that there is not a one-size fits all model/approach that will work for all bot detection requirements. By merging and expanding on past bot detection research, we can create an easy to use "tool box" that can address several bot-detection requirements. The evaluation provided later in this paper will demonstrate that key models in the bot-hunter suite of tools are equivalent or better than state of the art models.

3.3 Data

Our team used the Twitter REST and Streaming API's to access the data used in this research effort. Details of this process are provided below.

3.3.1 Overview of Available Data

Research is loosely divided between account-focused data collection strategies and topical or stream-based collection strategies. Account-based approaches will only use data objects directly tied to the user (user JSON object, user time-line object, etc). Stream-based approaches extract features from a given topical stream or twitter stream sample. These stream based features are often network features, but represent a small fraction of the ego-centered network of a given account. Our research therefore pursues an account-based approach to build a fuller representation of the account's ego network.

Researchers must find a balance between speed and richness of data. Past account-focused research generally falls into four *tiers*. Table 3.1 provides a description for each *tier* of data collection, the estimated time it would require collecting this data for 250 accounts, and the amount of data that would be available for feature engineering per account.

3.3.2 Data required for account conversation networks

Detailed ego network modeling of a Twitter account's social interactions requires Tier 3 data collection, but to date our team has not found any research that has conducted that level of data collection to model the network structures and social conversations that an automated Twitter account interacts with. In fact, few teams go beyond basic in-degree (follower count) and out-degree (friend count) network metrics found in Tier 1 meta-data. The closest effort to date is the *Botometer* model, which arguably operates at *Tier 2*. By adding the user timeline, *Tier 2* provides limited network dynamics, to include being able to model hashtag and URL co-mentions in a meta-network (see Figure 3.1). The resulting *timeline* based network, however, lacks comprehensive links between alters. While the time-line can provide rich temporal patterns, we found that it lacked sufficient structure to model the ego network of an actor.

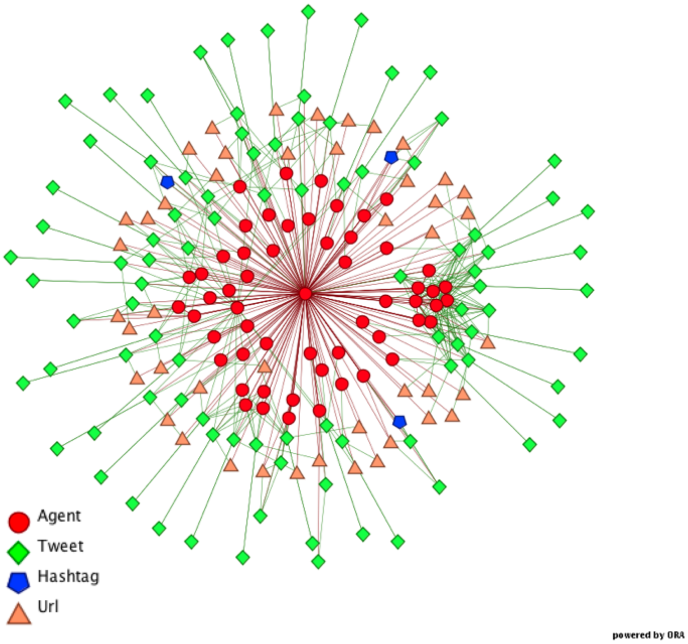


Figure 3.1: Leveraging only user *timeline* provides limited network features in a *star graph*

We set about to build the social network and social conversations that a twitter account is interacting with. We also tried to do this in a way that would expedite the time it takes to collect the data and measure network metrics. Our initial goal was to collect data, build the feature space, and classify an account within 5 minutes. We selected the five minute limit in an attempt to process ~ 250 accounts per day with a single thread

To collect the necessary data, we executed the following steps sequentially:

1. Collect user data object
2. Collect user timeline (last 200 tweets)
3. Collect user followers (if more than 250, return random sample of 250 followers)
4. Collect follower timelines (last 200 tweets)

When complete, this data collection process (illustrated in Figure 3.2) creates up to 50,000 events (tweets) that represent the conversation and virtual social interaction that the user and their followers participate in.

The resulting network, while partially built on social network structure (the initial *following* relationship), is primarily focused on the larger conversation they participate in. We initiated the single seed snowball by querying *followers* rather than *friends* since *followers* are much less controlled by the bot-herder, and contain fewer news and celebrity accounts. We conducted a *timeline* rather than *followers* search for the 2nd hop of the snowball to overcome rate-limiting constraints and to model the conversation network rather than directly model the social network. This single seed snowball process conducts a limited breadth-first-search starting with a single seed and terminating at a depth of 2.

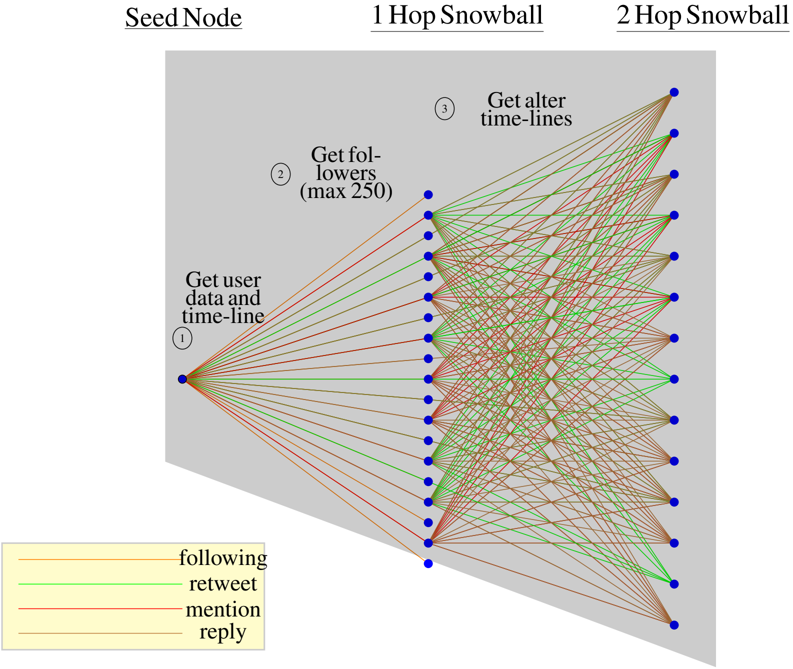


Figure 3.2: Illustration of 2-hop Snowball Sampling: Conversation of Target Node and Followers. First get followers of target node (if more than 250, sample followers). Then get timelines of alters. Use timelines to draw connections to accounts that alters *retweet*, *reply*, and *mention*

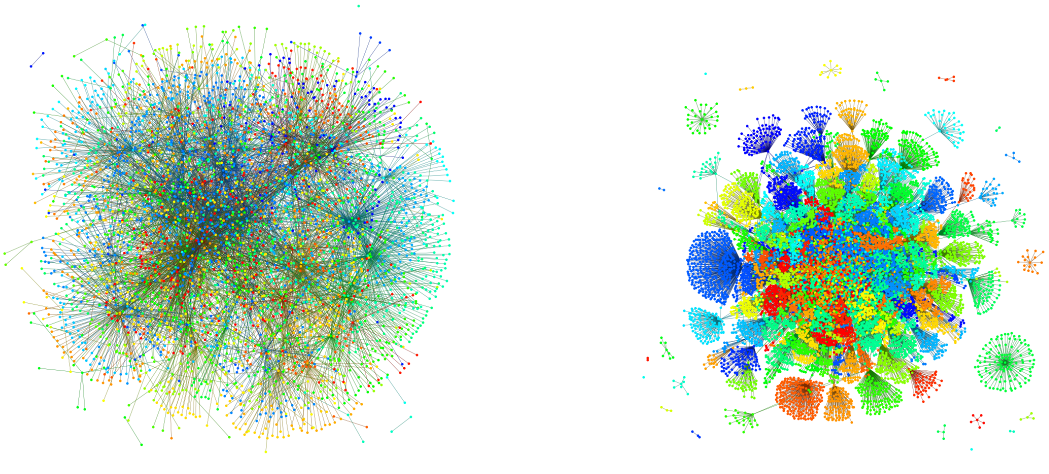
Artificially constraining the max number of alters at 250 was a modeling compromise that facilitates the self-imposed 5 minute collect/model time horizon. The choice of 250 allows our process to stay under 5 minutes, and also represents the upper bound of Dunbar's number (the number of individuals that one person could follow based on extrapolations of neocortex size) [86]. Additionally, in evaluating a sample of 22 million twitter accounts, we found that 46.6%

had less than 250 followers. This means that approximately 50% of accounts will have their entire ego network modeled. Bots tend to have fewer followers than human accounts and from the 297,061 annotated bot accounts that we had available for this research, 72.5% of them had fewer than 250 followers. Given that this compromise will only affect 25% of the bot accounts and 50% of all accounts, we felt that it was appropriate.

We used this data to create an agent to agent network where links represent one of the following relationships: mention, reply, retweet. These collectively represent the paths of information and dialogue in the twitter "conversation". We intentionally did not add the *follow/friend* relationships in the network (collected in the first hop of the snowball) since *follow/friend* relationships are an easy metric for bot herders to simulate and manipulate with elaborate bot nets. Complex conversations, however, are much harder to simulate, even in a virtual world. Additionally, adding the *following* links between the ego and alters would have created a single large connected graph. By leaving them out, we were able to easily identify the natural fragmentation of the social interaction.

3.3.3 Visualizing conversations

During our initial exploration, we visualized these *conversations* for both human accounts and bot accounts. A comparison of these conversations is provided in Figure 3.3. Note that bots tend to get involved in isolated conversations, and the followers of the bot are very loosely connected. The network created from a human virtual interaction on Twitter, is highly connected due to shared friendship, shared interests, and shared experiences in the real world.



(a) Human *conversation*

(b) Bot *conversation*

Figure 3.3: Differences between a human Twitter conversation(s) and a bot Twitter interactions (networks colored by louvain group) [33]

3.3.4 Annotated Data

For annotated bot data, we combined several legacy annotated bot data sets as well as some that our team has annotated during the development of the *bot-hunter* toolbox. Note that a Tier 3 model requires additional collection of friends, followers, and followers timeline, and therefore requires accounts that are not suspended. Several rich annotated bot data sets were used for our Tier 1 and Tier 2 models have a high number of suspended accounts and, therefore, were not used for the development of a Tier 3 model. These datasets will still be discussed in the results and evaluation sections since they were used in the development of Tier 1 and Tier 2 models.

The first data set used for Tier 3 training data is a large diverse bot data set that was annotated by detecting 15 digit random alpha-numeric strings as indicated in [31] (a data annotation method using a Tier 0 model). This method provided 1.7 million annotated bot accounts. From this data we built network metrics on 6,874 of these accounts. The second data set is from the Debot bot detection system [61] which includes bots that were found due to correlated activity. Using the Debot API, our team extracted 6,949 of these accounts, from which we built network metrics on 5,939 accounts. Additionally, we used the bot data manually annotated by Cresci et al in 2015 [70] and again in 2017 [71].

In the results section we will discuss several other data sets that were used to train our Tier 1 and Tier 2 models. These include the annotated data our team captured in a bot attack on the NATO and the Digital Forensic Labs [30]. This data will be referred to as NATO in the results. We also used the suspended Russian bot data set that Twitter released in October 2018 [236]. This data set primarily contains bot/cyborg/troll activity generated by the Russian Internet Research Agency (IRA) during the 2016 US National Elections. In our results sections, this data set is referred to as the IRA data. Finally, we used a large data set of suspended accounts. To acquire this data, our team streamed the 1% Twitter Sample for 7 months, and then went back to discover which of the accounts had been suspended. A similar data collection technique was used by Thomas et al. in [230].

The IRA and *suspended* data sets were only used for Tier 1, since timeline and followers were not available for Tier 2 and tier 3. For the NATO accounts, 96% of the accounts in this dataset have been suspended. We were able to collect sufficient data for Tier 2, but not Tier 3. A summary of each data set is provided in Table 3.3 and cross-walked with the models that it was used with. Note that the Varol data set is not provided here and was not used in our latest bot-hunter models since it is dated and did not perform well.

These data sets contain a wide variety of bots. The Varol data set was founded on the original 2011 Caverlee [150] Honey Pot data, but was supplemented with manual annotations (we leveraged only the manually annotated data). The Cresci data contains both traditional spambots (largely commercial spambots) as well as social spambots (both commercial and political). The random string data contains a large variety of bots ranging from political bots focused on the Middle East to hobby bots focused on Japanese Anime. The Debot data is also fairly diverse, with the one unifying feature that they all have content and timing correlated with other accounts. The differences in these bots are demonstrated in the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction that we conducted on 2000 randomly sampled accounts from the combined data set (see Figure 3.4). Here we see that the Debot data appears to be separate and different from the Varol, Cresci, and Random String data, which appear to be more

Table 3.3: Data Description

Training Data	Description	Tier1	Tier2	Tier3
Cresci 2017	Manually annotated by Cresci et al in 2017 [71]	X	X	X
Cresci 2015	Manually annotated by Cresci et al in 2015 [70]	X	X	X
Debot data	Accounts labeled as bots by the Debot bot detection system [61]	X	X	X
NATO	Data our team captured in a bot attack on the Digital Forensic Labs and NATO [30]	X	X	
Suspended Accounts	These are accounts that were suspended by Twitter	X		
Random String Accounts	Accounts with 15 digit random alpha-numeric strings as screen names [31]	X	X	X
IRA Data	Suspended Russian bot dataset that Twitter released in October 2018	X		
Combined Data	Combination of data listed above	X		

uniformly distributed in this 2-dimensional representation of the data.

In order to train a model, we also needed accounts annotated as *human*. For earlier Bot Hunter models, we used the Twitter Streaming API to collect a sample of *normal* Twitter data, intentionally collecting both weekend and weekday data. This provided 149,372 accounts to tag as *human* Twitter accounts. Of these accounts, we were able to collect/measure network metrics on 7,614 accounts. Past research has estimated that 5-8% of twitter accounts are automated [239]. If this is true, then we mis-labeled a small amount of our accounts as *human*. In earlier models we found that this was acceptable noise in the data, and created models that matched or exceeded the state of the art and were found useful in classifying bots.

By the time we reevaluated this technique in 2020, we found that while bots still account for a low percentage of Twitter accounts, they account for a much higher percentage of tweets produced. By sampling from the 1% Twitter Sample, we were getting a much higher percentage of bot accounts that were being mis-labeled as human. To test this we sampled 1 million Tweets from the 1% sample on 19 March 2020, and found that 65% of the accounts had bot characteristics (bot probability greater than 0.5). This indicates that sampling users from the 1% sample no longer provide a high concentration of human accounts.

In order to get past this limitation, for Bot Hunter models produced in 2020 and beyond we attempt a sampling technique that focuses on dense 'human' networks. By starting with human seed accounts that are highly likely be followed by other humans, we are able to conduct a 1 or 2 hop snowball sample along following links to get likely human accounts. For example, we

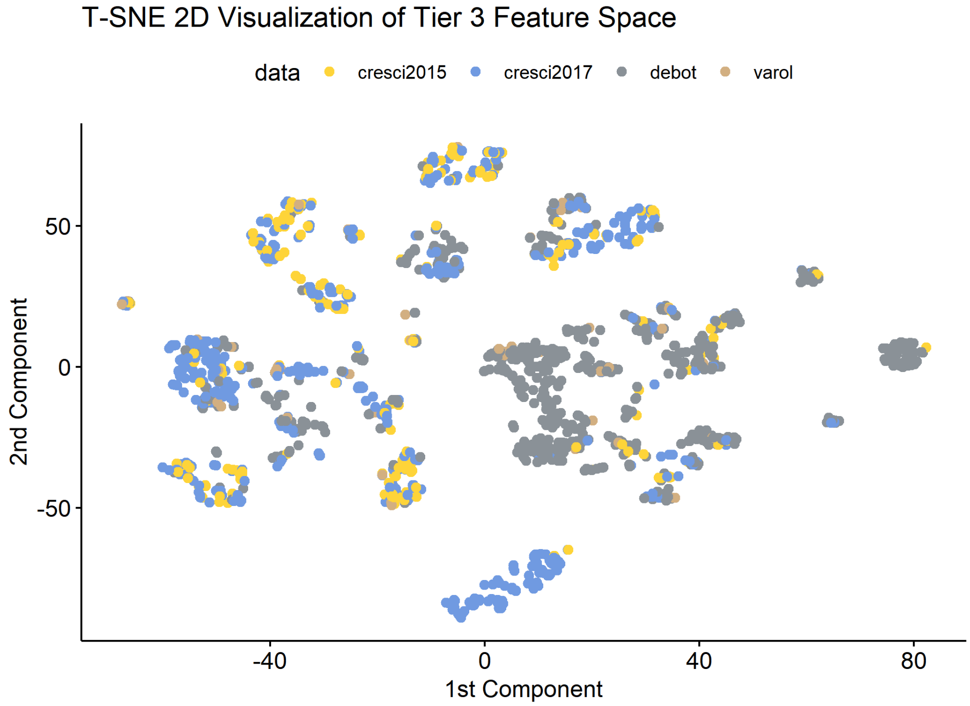


Figure 3.4: t-SNE Dimentsionality Reduction of Tier 3 Feature Space (by bot dataset)

are able to choose certain business or educational users that we identify as highly likely to only have following links with other human users, and use them as seed nodes. To make sure that we fully represent the diverse non-bot accounts, we need to then add multiple other types of non-bot accounts. For our purposes, we added celebrity, commercial, media, and government accounts so that we have a diverse set of 'non-bot' accounts. This technique provided better models for the emerging bots in 2020 and beyond.

3.4 Feature Engineering

In this section we will introduce our feature engineering for user, content, temporal, and network features. We extracted features from Tier 0 through Tier 3, with a focus on measuring the importance of features extracted from Tier 3. The table of proposed features is provided in Table 3.4. All new features (beyond the features we presented in [33]) are in bold, and from our research most of these have not been used with an ego-network collected with snowball sampling.

Note that our tiered approach is cumulative, meaning Tier 3 feature space includes features from Tier 0, Tier 1, and Tier 2. The Tier 3 model therefore includes the Tier 2 network features created by building an entity (mention, hashtag, and URL) co-mention network based only on the user's time-line (last 200 tweets). These Tier 2 network features are distinguished in our results section by the *entity* prefix.

We hypothesize that the network metrics for human conversations will have different distributions than those made by bot accounts. We also believe that these differences would provide increased performance in traditional machine detection algorithms.

We have not found research that has built a snowball sampling network for bot detection, and believe that all of the Snowball Sampling ego network features in our model are novel. To collect these at scale, our team built a Python package that wrapped around the *networkx* package [118]. We leveraged known network metrics, which are provided in Table 3.4 with references.

3.4.1 Network Features

We constructed an ego network from the data collected from snowball sampling, extracting metrics from this network to develop robust features for bot detection. As discussed earlier, this network consisted of the conversation of the account in question and up to 250 of their followers. All nodes were Twitter accounts, and links were means of directed communication in the Twitter ecosystem (retweet, mention, reply). From this network we developed basic network metrics, component level statistics, centrality metrics, triadic relationship metrics, and clustering related metrics. The basic network metrics are widely used and listed in Table 3.4. The other categories of metrics are described below.

Given that we did not include the *following* link in our network construction, these networks were not fully connected. As seen in Figure 3.5, information from these disconnected components could be valuable in distinguishing real human networks from networks dominated by bots. Our features therefore contain multiple metrics measuring number and size of network components.

We included several network centrality metrics in our feature space, and found that they were routinely strong bot predictors. These metrics included mean degree centrality, mean eigenvector centrality, and mean K-betweenness centrality where $K = min(500, N_{nodes})$.

In addition to analyzing the components, we also computed Louvain grouping and developed metrics based on these groups. We chose the Louvain grouping algorithm given its proven performance on larger data sets. Having computed the Louvain groups, we included metrics such as number of groups and size of largest Louvain group.

Given the importance of triadic relationships in social networks discussed above, we have included several features based on these relationships. These include a full triadic census, number of Simmelian ties, and the clustering coefficient. Calculation of Simmelian ties [142] was not available in the *networkX* package. Our team therefore created a Python implementation of Dekker's version [81] of the original algorithm [142].

Table 3.4: Features by Data Collection Tier (New features not presented in [33] highlighted in bold)

Source	User Attributes	Network Attributes	Content	Timing
User Object (Tier 1)	screen name length default profile image? entropy screen name has location? total tweets source (binned)	number of friends number of followers number of favorites	Is last status retweet? same language? hashtags last status mentions last status last status sensitive? ‘bot’ reference?	account age avg tweets per day
Timeline (Tier 2)		number nodes of E number edges density components largest compo degree/between centrality	mean/max mentions mean/max hash number of languages fraction retweets	entropy of inter-arrival max tweets per hour max tweets per day max tweets per month
Snowball Sample (Tier 3)	**% w/ default image** **median # tweets** **mean age** **% w/ description** **% many likes & few followers**	# of bot friends number of nodes number of links density number of isolates number of dyad isolates number of triad isolates number of components > 4 clustering coefficient transitivity reciprocity degree centrality K-betweenness centrality mean eigen centrality number of Simmelian ties number of Louvain groups size of largest Louvain group ego effective size full triadic census median followers median friends	**# of languages** **mean emoji per tweet** **mean mention per tweet** **mean hash per tweet** **% retweets** **mean jaccard similarity** **mean cosine similarity**	**mean tweets/min** **mean tweets/hour** **mean tweets/day** **% don’t sleep**

3.4.2 Content Features

We felt we could leverage the large amount of content available from the snowball sample to develop predictive features. This was not done in [33], and was added in a recent version of the *bot-hunter* framework.

These features include the number of languages used in the network, as well as some key summary statistics on entities, including mean emojis, mentions, and hashtags per tweet, as well as the percentage of retweets.

We also wanted to have several measures of similarity of text between the various communicators in the network. This search for similarity measures was motivated by the fact that many bot networks post very similar or conversely very diverse content, and we felt that these measures of similarity may be distinguishing.

To compute similarity, tweet content in the network was aggregated by user. Once aggregated, the content was cleaned and parsed (cleaning included conversion to lower case and removal of punctuation). We did not remove stopwords. The parsed data was then converted to a document term matrix with raw counts (we chose not to normalize the data since the variance on tweet length is artificially constrained to 280 characters). The document term matrix was then used to compute both the Jaccard and Cosine Similarity, which were used as features.

3.4.3 User Features

The newest version of the Tier 3 classifier also includes several aggregate user attributes that were not leveraged in earlier versions. While many of these are self explanatory, we did want to describe two novel metrics that have not been used before.

Recently, several experts in online disinformation have highlighted how recent online bots seem to produce tweets that are far more popular than the account itself [182]. This phenomena is the result of accounts in large bot-nets that create messages that are then *pushed* by the entire network, resulting in reach that far exceeds expectations given its modest beginning.

To find this phenomena, we devised a simple heuristic that determines if any original (non-retweet) tweet is more popular than its account. This heuristic is defined as:

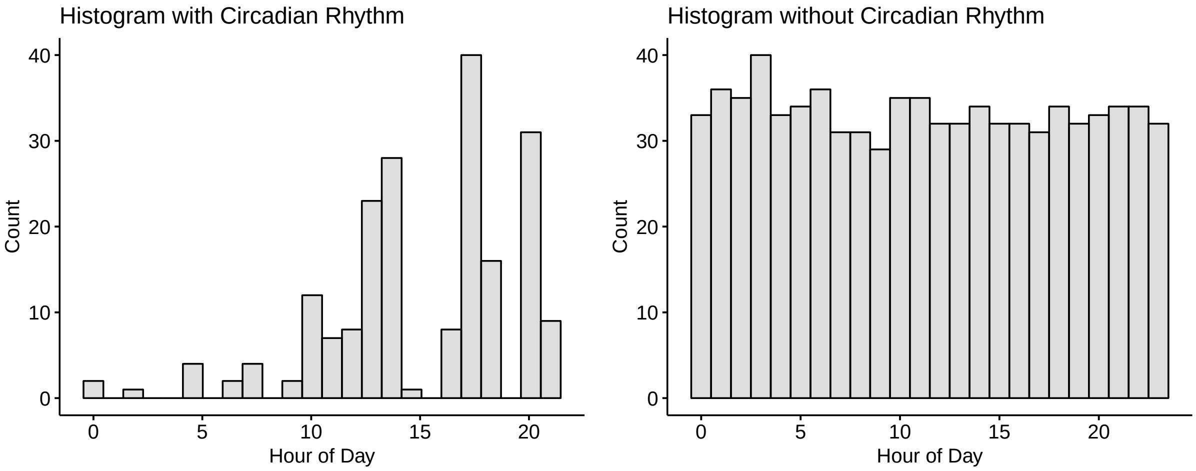
$$P_{user} = retweets > 2 \times max(followers, friends)$$

where the boolean measure for a user is defined as $True$ if any tweet receives 2 time more retweets than the highest value of its in-degree or out-degree. This metric is leveraged in two new features, one at the user level (Tier 2) and one at the Network Level (Tier 3). The user level flags the user if any tweet is flagged as True, and the network metric measures the fraction of tweets produced by the network that are flagged by this heuristic.

3.4.4 Timing Features

Like user features, most of the temporal features listed in Table 3.4 are self explanatory. We did develop a heuristic method that measures whether an account has daily rhythms. Most human users will have surges in activity based on their daily routines, and will have a measurable drop in activity that aligns with their sleep activity. Bots, on the other hand, do not require these circadian rhythms, and some bots are programmed to produce content spread uniformly across the hours of the day. We developed the heuristic described below to flag these accounts.

To measure whether an account has human circadian rhythms, we first aggregate their tweets by hour of day after ensuring that the account has produced enough data (at least 50 tweets). Given there is enough data, we next determine whether this hourly distribution is uniformly distributed by normalizing it and conducting the Kolmogorov-Smirnov non-parametric test for uniformity. A p-value greater than 0.5 provides strong evidence of non-human circadian rhythms.



(a) Human *circadian rhythms*

(b) Bot *without* circadian rhythms

Figure 3.5: Differences between a human and bot 24-hour circadian rhythms

It is important to note that, while some bots exhibit this lack of circadian rhythm, it only takes a few lines of code for a bot manipulator to give a more realistic temporal pattern. Nonetheless, this remains a strong indicator of bot activity.

3.5 Modeling

As indicated above, all feature engineering was conducted in Python using several custom Python packages that were developed for the *bot-hunter* framework. These packages build the feature space for Tier 1, Tier2, and Tier 3 models, which is then trained using the steps outlined below.

For training all data sets, human data was sampled so that the classes were balanced. The random forest algorithm was used because of its superior performance on Tier 1 data [30] and its use in other bot detection algorithms [239]. In Table 3.5 we revisit model comparison in order to verify that a random forest model is still appropriate for Tier 3 feature space. We see that random forest still provides superior performance, and in general is not as computationally expensive as some of the other models. Training, evaluation, and testing were conducted in the *scikit-learn* Python package [189]. Tuning of the Random Forest algorithm was conducted through random search of parameter options while using 3-fold cross-validation.

Table 3.5: Comparing Algorithms for Tier 3 Bot Detection

Model	**Accuracy**	**Precision**	**Recall**	**AUC**	**F1**
Naïve Bayes	0.562	0.541	0.864	0.563	0.665
Decision Tree	0.950	0.949	0.952	0.950	0.951
SVM	0.952	0.969	0.933	0.952	0.952
Logistic Regression	0.951	0.940	0.965	0.983	0.952
Random Forrest	0.955	0.955	0.956	0.986	0.956

The *bot-hunter* behavior returns both a binary classification and an estimate of probability. The estimate of probability is provided by the Random Forest classifier by measuring the proportion of votes by trees in the ensemble. The binary classification result is evaluated by classifying accounts based on a probability threshold of 0.5. The binary classification feature of the results allows researchers to have a consistent threshold to compare results, while the probability allows users to tune a threshold for a given use case.

3.6 Results

After building the network metrics for all bot data sets, as well as the annotated *human* data, we built and evaluated Random Forest models for each of the data sets. Training, evaluating, and testing were conducted at Tier 1, Tier 2, and Tier 3 where possible. We evaluated in-sample performance with 10-fold cross-validation measuring multiple evaluation metrics, which are provided in Table C.3 and Figure 3.6.

From the results presented in Table C.3 and Figure 3.6, we see that Tier 1 models continue to provide solid performance, even with basic features extracted from the user profile and last status. We also observe improvement between Tier 1 and Tier 2 and between Tier2 and Tier 3 for all models. Using a combined data model we found that the Tier 2 improvement over Tier 1 is statistically significant ($p - value = 1.303e - 10$), as is the Tier 3 improvement over Tier 2 ($p - value = 1.101e - 06$). In Figure 3.6 we also see that the Random, NATO, and IRA data provide the highest in sample cross validation performance, while models trained on Debot Data and Suspended data offer lower cross validation performance. This likely indicates a wider variety of bot types in the the Debot and Suspended data.

Table 3.6: Table of Results for Combined Data (Tier 3)

Tier	Accuracy	F1	Precision	Recall	ROC AUC
Tier 1	0.7964	0.7729	0.8677	0.6969	0.8680
Tier 2	0.8335	0.8181	0.8970	0.7522	0.9179
Tier 3	0.8577	0.8478	0.9042	0.7983	0.9410

Further, in Figure 3.7 we see the top features for all Tier 1, Tier 2, and Tier 3 models in the *bot-hunter* suite of tools. These figures represent the percentage that each feature contributed to the model predictions. We see that network features provide strong features in the model.

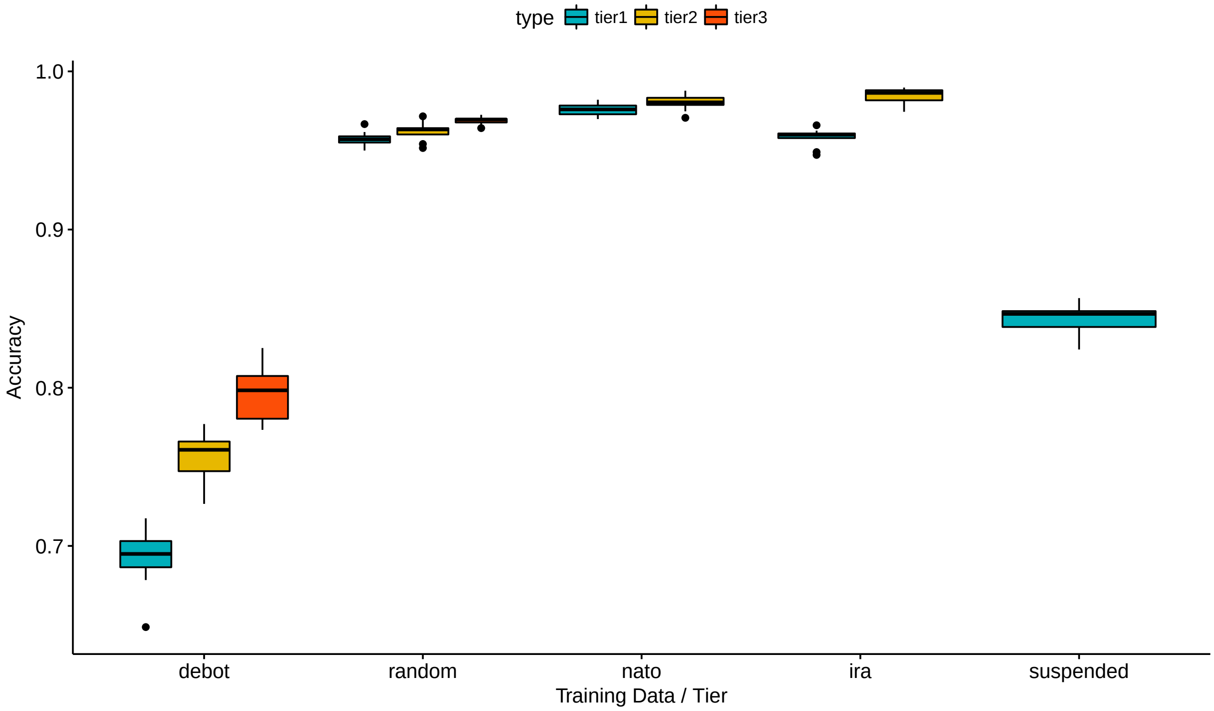


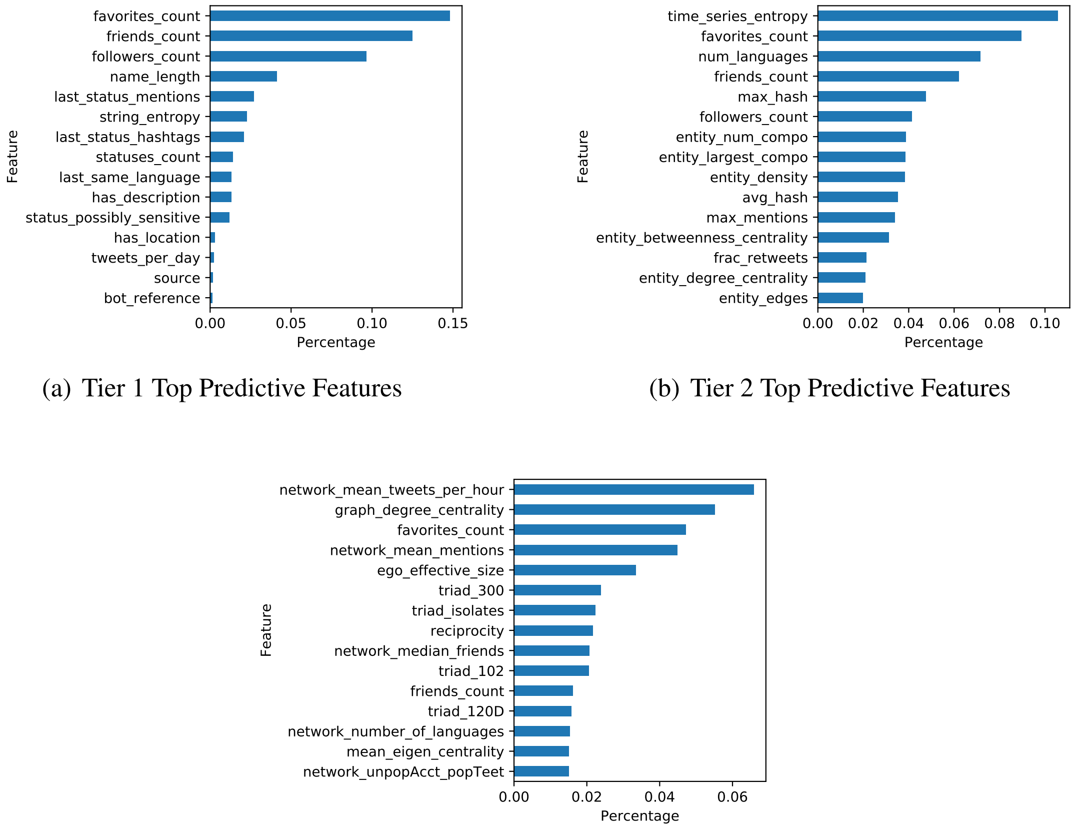
Figure 3.6: Results by training data and by *Tier*

This demonstrates that these values, while tedious to collect, transform, and model, provide strong predictive features that are difficult for bot *puppet master's* to manipulate. In these data sets network centrality, network connection, network timing, and network content all provide predictive value.

3.7 Evaluating Against State of the Art

Given that this is the last Tier of the *bot-hunter* suite of tools, we wanted to evaluate the models as well as various training data that is available. We also wanted to compare the models in the *bot-hunter* suite of tools to existing models, namely the Botometer and Debot models. To do this, we set out to find a test that wasn't biased toward any given model, meaning the test data could not be derived from the training data of any of the models being compared.

To find an unbiased data set, we manually annotated 337 bot accounts. To do this, we started by manually finding several seed bots related to the Swedish elections, separate Russian propaganda bots, and bots found in Middle East conversations. We then manually snowballed out on the followers and followers of followers, manually identifying additional bots. In this evaluation we leveraged the visualizations and metrics provided in the TruthNest tool to aid in making our determination. The TruthNest Tool originally was an EU-funded Reveal project developed to evaluate Twitter accounts for automated activity. While this tool was not evaluated in our test, it was used to assist in labeling bot accounts. TruthNest has instituted a paywall since our use of it. Human users were sampled from the Twitter stream and manually verified. The test data was



(a) Tier 1 Top Predictive Features

(b) Tier 2 Top Predictive Features

(c) Tier 3 Top Predictive Features

Figure 3.7: Comparison of Top Features for all three Tiers of *bot-hunter*

balanced (337 bots, 337 users).

In evaluating our Tier1, Tier2, and Tier3 models, we also wanted to evaluate which training data and model combination generalizes to new data. Our models were trained on the data and at the tiers described in Table 3.1. All *bot-hunter* and Botometer thresholds were set at 0.5. F1 performance for all models is provided in Figure 3.8 and detailed results are provided in Table 3.7.

In these results we first see Botometer demonstrates consistent solid performance in predicting new bots across all metrics. Note that Botometer typically has high precision but relatively low recall, resulting in high accuracy but somewhat lower F1 score. This means that Botometer is generally correct when predicting an account as a bot or human, but it fails to find a large portion of the bots that are present. The Debot algorithm provides high precision but extremely low recall, resulting in a low F1 score overall. The value of the Debot algorithm may indirectly lie in the data that it produces. Note that *bot-hunter* algorithms trained on Debot data performed well at all three Tiers, meaning that the Debot algorithm for finding correlated accounts produces great labeled data for other supervised bot detection endeavors.

For the bot-hunter family of models, we see that Tier 1 consistently performs well and seems

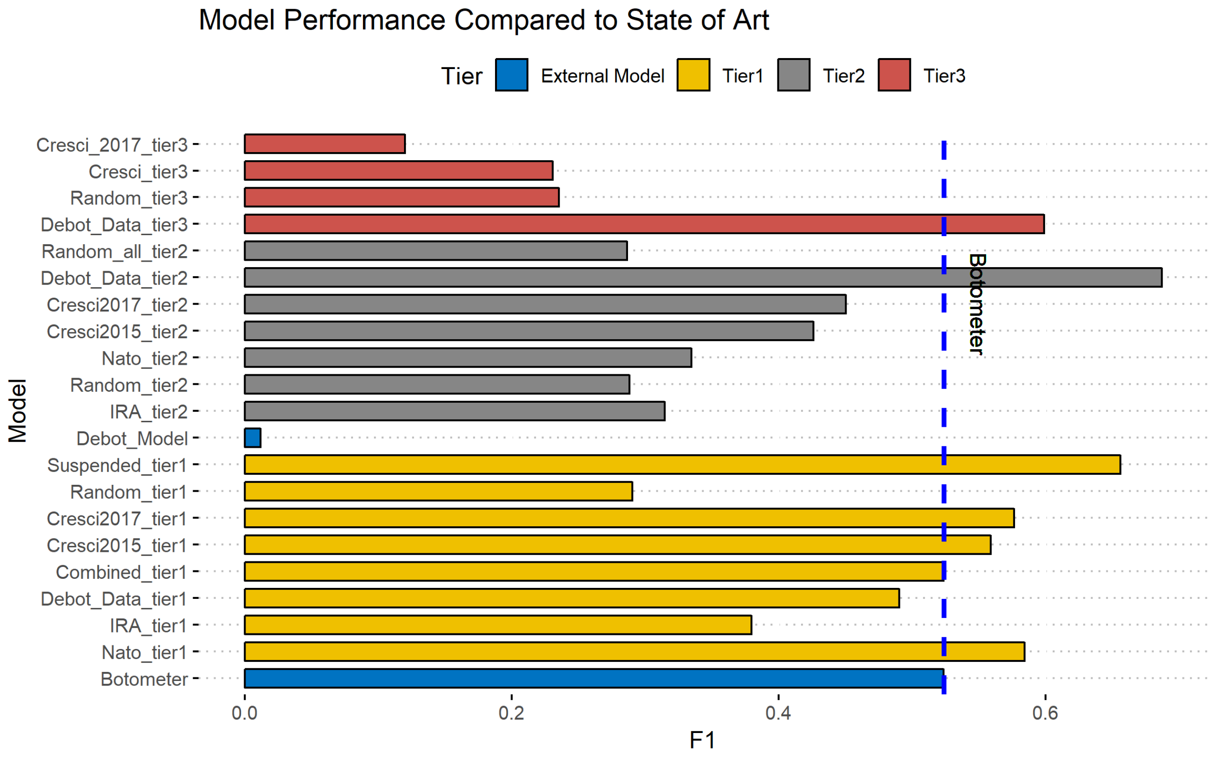


Figure 3.8: Results by training data and by *Tier*

to generalize to new data better than Tier 2 and Tier 3. Tier 2 still has high performance, given its ability to identify anomalies in content and in temporal statistics. Across the data sets, Tier 2 has a higher mean Accuracy and ROC AUC than Tier 1. Tier 3 has very high precision but low recall. It therefore produces predictions that are more reliable, but fails to find a large portion of the bots in the data. However, this model may become increasingly important in identifying sophisticated emerging bots.

As we look at the various training data used for training these models, we see that the models trained on suspended accounts or on data produced by the Debot model had the highest performance. As indicated earlier, this is likely due to these data sets containing a wide variety of bot "genres." We also see that the NATO data captured in the deliberate attack against NATO and the DFR labs continues to provide strong performance across all metrics. We found that few of the annotated data sets released by other researchers provided strong performance, especially when considering accuracy and ROC-AUC metrics. The Cresci data (both 2015 and 2017) appears to have high recall but low precision, with many false negatives. The models trained on the random string data also have low accuracy and ROC-AUC metrics, in this case caused by high precision but low recall. These random string accounts probably represent a limited band in the spectrum of bot types, and therefore do not generalize well to new data and different bot types.

The Venn Diagram of predicted bots is provided in Figure 3.9(a). This diagram shows the overlap of the predicted bots, but does not provide any information on predicted humans. We see significant overlap for all three models. We also notice that the Tier 2 model predicted the most accounts (330 accounts), while Tier 1 predicted 260 accounts and Tier 3 predicted 183 bot

Table 3.7: Detailed Results by Tier and Training Data

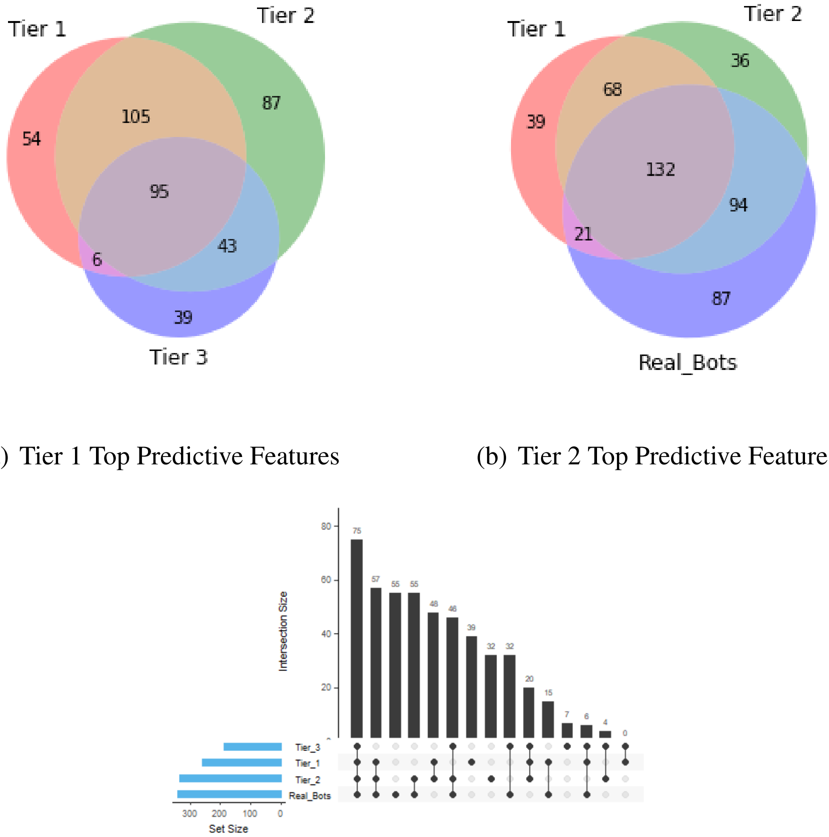
Tier	Training Data	F1	Accuracy	Precision	Recall	ROC-AUC	TN	FP	FN	TP
Botometer Model		0.524	0.657	0.858	0.377	0.587	256	55	200	108
Debot Model		0.012	0.502	1.000	0.006	0.503	336	0	335	2
Tier1	NATO	0.584	0.634	0.678	0.513	0.635	254	82	164	173
Tier1	IRA	0.380	0.597	0.830	0.246	0.598	319	17	254	83
Tier1	Combined	0.524	0.657	0.858	0.377	0.657	315	21	210	127
Tier1	Cresci2015	0.559	0.404	0.444	0.754	0.404	18	318	83	254
Tier1	Cresci2017	0.576	0.419	0.454	0.789	0.418	16	320	71	266
Tier1	Debot	0.490	0.527	0.533	0.454	0.528	202	134	184	153
Tier1	Random	0.291	0.572	0.855	0.175	0.573	326	10	278	59
Tier1	Suspended	0.656	0.713	0.821	0.546	0.713	296	40	153	184
Tier2	IRA	0.315	0.567	0.903	0.191	0.584	305	7	276	65
Tier2	NATO	0.335	0.574	0.909	0.205	0.591	305	7	271	70
Tier2	Cresci2015	0.426	0.596	0.824	0.287	0.610	291	21	243	98
Tier2	Cresci2017	0.451	0.600	0.799	0.314	0.614	285	27	234	107
Tier2	Debot	0.687	0.675	0.691	0.683	0.675	208	104	108	233
Tier2	Random	0.286	0.550	0.831	0.173	0.567	300	12	282	59
Tier3	Debot	0.599	0.674	0.837	0.466	0.683	281	31	182	159
Tier3	Random	0.236	0.533	0.810	0.138	0.551	301	11	294	47
Tier3	Cresci2015	0.231	0.541	0.918	0.132	0.560	308	4	296	45
Tier3	Cresci2017	0.120	0.507	0.880	0.065	0.527	309	3	319	22

accounts. The 95 accounts in the intersection contain 20 false positives (78.9% precision).

The Venn Diagram of predicted bots for Tier 1 and 2 compared to the *real* labeled bots is provided in Figure 3.9(b). This shows that Tier 2 is adding something to Tier 1, finding 94 additional accounts while only missing 21 of the accounts that Tier 1 found.

Figure 3.9(c) provides an *upset* visualization to fully explore the intersection of sets. This visualization demonstrates that our largest intersection is the intersection of all four sets. We also see in the upset graph the Tier 1 and in particular Tier 2 is important to the prediction success, though Tier 3 is also able to find 32 accounts that neither Tier 1 or 2 could find. These visualizations illustrate the importance of having a tool-box of models that can be used for predicting bots in any given scenario.

While we believe this evaluation is informative, there are several limitations in our evaluation method. We acknowledge that we were not able to completely remove bias, given that the mental heuristics we used to manually annotate accounts may have unintentionally mirrored the bot-hunter algorithms. Additionally, we acknowledge that the test set is still modest in size and, while somewhat diverse, does not represent the full spectrum of bot types. Finally, we acknowledge that any given model may perform better if the threshold is tuned for a given data set. Even with these limitations, we believe this test and evaluation is informative for our team and for the greater community.



(a) Tier 1 Top Predictive Features

(b) Tier 2 Top Predictive Features

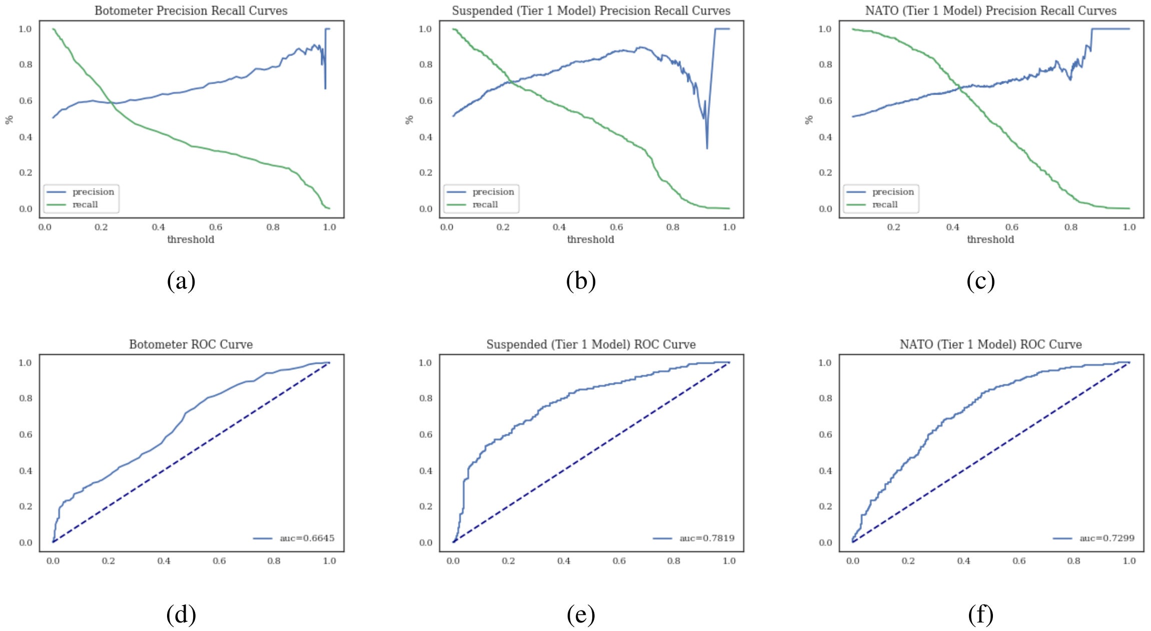
(c) Tier 3 Top Predictive Features

Figure 3.9: Upset Plot with Predicted Bots (Tier 1, 2, & 3) and Real Labeled Bots

3.7.1 Evaluating bot classification thresholds

The random forest model used in the bot-hunter suite of tools (and Botometer) provides a probability estimate rather than just a label. This allows researchers to estimate how strong a given prediction is. Every use case will require the analyst to determine the best threshold for establishing whether or not an account is likely a bot. To evaluate the best threshold for a given data set, a research team should explore several thresholds, each time sampling 50-100 accounts and manually labeling them to estimate a rate of true/false positives, true/false negatives. If possible they should attempt to construct a precision recall curve and/or ROC Curve, as demonstrated in Figure 3.10 using the Suspended, NATO, and Botometer models. Note that recall is always monotonically decreasing, but precision is not required to monotonically increase.

As seen in Figure 3.10, we generally recommend bot-hunter thresholds between 0.6 and 0.8. The exact choice in this range will need to be made by the research team, and is dependent on the data as well as the team's prioritization of precision vs. recall.



(a) (b) (c)

(d) (e) (f)

Figure 3.10: Using Precision-Recall Curves and ROC Curves to determine threshold

3.8 Applying Bot Detection to Swedish Election

Having completed the bot-hunter suite of tools, we wanted to leverage this toolbox in analyzing a stream of data from the 2018 Swedish elections. This is done as a case study to illustrate that bot-detection is not a “turn-key” solution, and also to provide practitioners with an example of an open source intelligence workflow.

Sweden held national elections on 9 September 2018 for its equivalent of a Parliament, known as the Riksdag. Swedish elections have historically lacked much drama or suspense, with the center-left Social Democrat Party dominating politics since 1914. In the 2018 election, however, their dominance was challenged by various nationalistic factions that capitalized on anti-immigrant sentiment.

Some of the political discourse surrounding the election transpired on Twitter, as seen in many recent national elections across the world. As this discourse grew, multiple researchers and news agencies saw rising disinformation and associated bot activity [224]. Simultaneously, the Swedish Defence Research Agency reported increased bot activity, primarily supporting right leaning, nationalistic, and anti-immigrant views [8].

As these bots grew in activity in this marketplace of beliefs and ideas, our team began collecting and analyzing streams from this discourse. To collect Twitter data around the Swedish National elections discourse, our team leveraged a spiral collection methodology, starting with content and geographic streaming, and then ‘spiraling’ into more thorough data collection around the important parts of the discussion. All collection was done through the Twitter Streaming and REST API’s using the Tweepy Python Package.

We started by identifying Swedish political hashtags through open source research, eventually identifying #svpol, #Val2018, #feministisktInitiativ, #migpol, #valet2018, #SD2018, #AfS2018, and #MEDval18. These hashtags were not selected because they cover the full spectrum of Swedish politics, but rather because there was open source reporting of some bot campaigns using these hashtags. We started collecting on these hashtags using both the Streaming and REST API's (the streaming API allows us to easily collect going forward while the REST API allows us to retroactively collect past data). Simultaneously we collected data that was 'geo-associated' with the Scandinavian peninsula, using a bounding-box search method.

As we began to collect content and geo-referenced data, we monitored other trending hashtags and added them to the collection query. After launching the exploratory data analysis discussed below, we would also collect user's friend and follower relationships as well as user's historical timelines for accounts of interest. This continual return to the Twitter API creates the spiral nature of our collection process.

For the Swedish Election Event we collected 661,317 tweets produced by 88,807 unique users. This creates a political *conversation* that contains 104,216 nodes and 404,244 links with a density of 0.000037.

For bot detection in the Swedish Election stream our team found that a 65% probability was appropriate. Given that we were performing this evaluation on 104,216 nodes, we used the Tier 1 model. This model is our best model for getting an accurate prediction on high volume of accounts.

Note that we usually conduct other data enhancement as well, including sentiment analysis with NetMapper as well as geo-inference based on [132]. All enrichments are made available in easy formats that allow tools to merge them with existing event data.

3.8.1 Exploratory Data Analysis

Our exploratory data analysis focuses on narratives, time, place, groups, and individuals. Our analysis typically starts with some type of temporal analysis. This allows us to see distributions over time. We try to look at overall temporal distribution, bot activity over time, as well as changing narratives over time.

Our exploration of content and narratives starts with analysis of words and hashtags across the entire corpus, and then we explore narratives associated with topic groups (these are groups that talk about the same thing but may not be connected in the social network or conversational network) and social network group (these are groups that are connected, but may not talk about the same thing). We leverage latent Dirichlet allocation [45] for topic group analysis, and content analysis by Louvain group [46] as a way to "triage" network groups. Table 3.8 provides the top 8 words by Louvain Group for the Swedish elections. In this we already start to see groups that are focused on immigration, particularly immigration from Muslim countries. We also see at least one group that is mixing conversation about religious beliefs with political discourse. Finally and just as important, "triaging" the data like this allows us to identify groups like Group 0 that don't appear to have any topics of interest.

Network analysis of groups and individuals is done almost exclusively in the ORA Network Analysis Tool. We typically start by visualizing a reduced conversational network. Nodes in this network represent Twitter accounts, and links represent a conversational action in the Twitter ecosystem (reply, retweet, mention). These networks are typically too large to visualize, so we

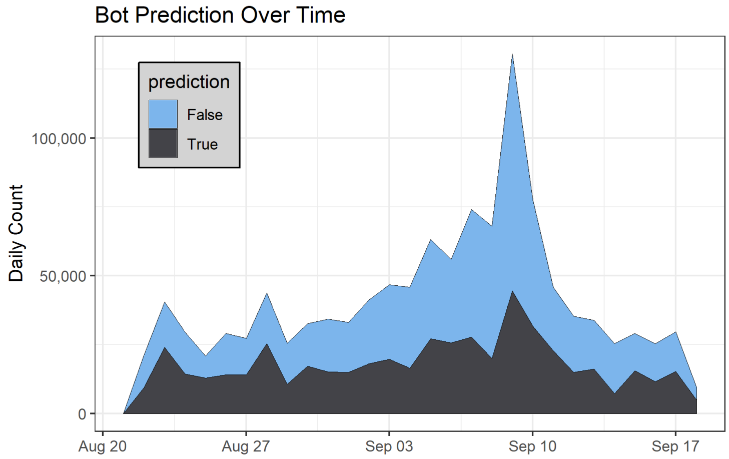
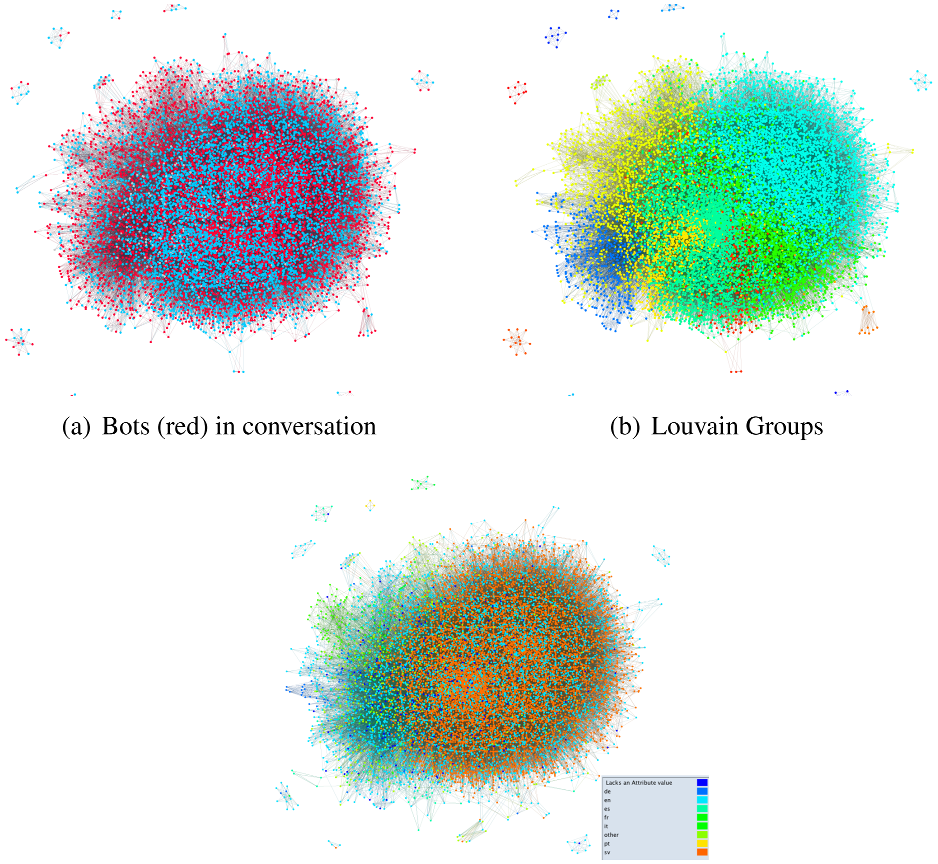


Figure 3.11: Bots as a proportion of total volume over time

Table 3.8: Content Analysis by Louvain Group

group	# Tweets	# Nodes	Top 8 Words By Louvain Group			
0	15,708	4,675	video 2018	gillade fortnite	lade world	spellista part
1	31,059	5,688	country n. . .	voters number	refugees capita	82 reported
2	102,146	14,538	sweden swedish	election results	epp left	sd poll
3	306,352	17,600	m6aubkudbg sverige	jesus gud	kristus namn	varnar fader
4	8,353	3,137	sweden amp	swedish vote	muslim democrats	election gang
5	40,585	9,110	sverige valet	sd jimmie	svenska år	åkesson svt
7	82,708	12,300	sd valet	sverige jimmie	rösta parti	åkesson val
8	17,675	4,000	sd politik	friend claeson	american tånkt	rösta frågar
9	7,144	5,217	sverige stefan	löfven kristersson	sd amp	moderaterna rösta
10	7,569	5,214	sverige afs	riks sweden	sd svenska	alternativ hahne

reduce the network by taking the K-core so that we have the core 15,000 to 20,000 nodes. Once this is done, we color the network by *bot* or *human*, by language, and by Louvain grouping (see Figure 3.12). This coloration helps us better understand the groups and their relationship to each other. Finally, we reduce the network to only include reciprocal links. This usually reduces the network significantly, and in Twitter provides the best proxy for a true social network.



(a) Bots (red) in conversation

(b) Louvain Groups

(c) Language Distribution in Network

Figure 3.12: Exploring the Twitter Conversational Network Surrounding online discourse on Swedish politics

We then explore the influential accounts and influential bots in the network. The ORA Network analysis tool provides several reports that analyze nodes by a variety of centrality measures, and assists translating their role in the network. For the Swedish network, we found several bots with high *betweenness*, indicating that these bots were influential in that they connect individuals and groups. With further exploration, it appeared that these bots, in connection with other accounts, were trying to bridge several communities with nationalistic and anti-immigrant groups and narratives.

We leverage the *bot-hunter* Tier 2 and Tier 3 models during this phase of analysis. As we identify influential accounts, we check them in a Tier 2/3 *bot-hunter* web application that allows us to thoroughly explore the account and conduct a more accurate Tier 2 or Tier 3 bot prediction.

Table 3.9: Bot detection F1, Precision, and Recall scores. All models but Botometer trained on Debot data. Top-2 F1 scores are emboldened, the state-of-the-art score is marked with an asterisk.

Model	F1	Precision	Recall
Botometer	0.524	0.858	0.377
Debot	0.012	1.00	0.006
Bot-Hunter Tier1	0.656	0.821	0.546
Bot-Hunter Tier2	**0.687**	0.691	0.683
Bot-Hunter Tier3	0.599	0.837	0.466
Graph-Hist	**0.740***	0.683	0.807

These applications also allow us to explore in-depth visualizations of the activity of the account.

Bot-detection is therefore a part of the overall open source intelligence workflow, trying to identify relevant information about how the world works to inform decision maker situational understanding and decisive action. In this case, our research validated research of large bot activity within the Swedish political discourse on Twitter and provided identification of narratives (primarily nationalistic and anti-immigrant, anti-Muslim, and some anti-Semitism). We were also able to identify influential accounts that were attempting to connect individuals and online communities with extremist content. This type of information informs leaders of current disinformation strategies allowing them to better prepare their government and their populace for similar disinformation campaigns in their country.

3.9 Alternate Tier 3 Approach

In [162] we propose an alternate Tier 3 model to the one discussed above. Rather than use graph metrics as a proxy for the graph, we use a graph neural network approach to classify the graph itself. In the alternate methodology, we use a technique we call Graph-Hist: an end-to-end architecture that extracts a graph's latent local features, bins nodes together along 1-D cross sections of the feature space, and classifies the graph based on this multi-channel histogram. While still computationally costly, we found that the neural network approach outperforms other approaches as seen in Table 3.9.

3.10 Retraining Bot Models

Bots are continually changing and adapting, meaning the bot detection models are ephemeral. As time marches on, the performance of supervised machine learning models slowly degrades until the model becomes obsolete. Several authors have discussed this phenomena [260]. This means that models must be incrementally retrained and validated. The retraining process will improve the training data and possibly the feature representation.

Teams that are maintaining a bot detection capability must constantly identify and collect new data on different genres of emerging bots. This includes collecting data on known bot events as well as trying to collect purchased or suspended accounts. This new data should be set aside for the next version of the bot model.

As the team prepares new bot data, they should also explore the accounts and their respective campaigns to ensure their current model features will capture differences between human and bot behavior. As new features are identified that may assist in distinguishing bots from human accounts, they can be added to the model feature set.

The frequency of model retraining must consider the trade-offs of model performance with model stability. The upgrade timing must be frequent enough to maintain predictive performance on emerging bots. However, it must also ensure that it is not so frequent as to disrupt on-going research projects. New model versions will create differences in prediction, which means that statistics produced by one model cannot be compared with statistics produced by a newer version. Our approach with Bot-Hunter is to upgrade approximately yearly, and to make the legacy model available for teams that are in the midst of a longitudinal investigation.

3.11 Extending Bot-Hunter to other Platforms

The Bot-Hunter feature space and data pipeline has been designed specifically for Twitter data. We chose to focus on Twitter because it increasingly hosts the global conversation, whereas Facebook and other platforms largely host local and topical conversations. That being said, we've been asked if the Bot-Hunter models and methodology could be extended to other platforms such as Facebook, YouTube, Instagram, Gab, Reddit, and others.

The concept and framework of supervised machine learning can be extended to any of these platforms, with varying degrees of performance. Some of the feature engineering can extend to other platforms, since these platforms also have account features, temporal features, and network features. While the features may be similar, their strength in predicting malicious accounts will likely vary among the platforms. Additionally, each platform will have unique limitations due to the data that is available through their respective Public API.

Additionally, other model approaches may work better on other platforms than Twitter. For example, multiple research efforts have successfully used unsupervised network based models to find malicious accounts on Facebook [177, 241, 262]. The strength of *friend* ties is stronger on Facebook than Twitter, and may be easier to extract (note that the *friend* and *following* data has the strictest rate limiting constraints on the Twitter Public REST API).

3.12 Conclusions

In our pursuit of a multi-model bot detection toolbox, this chapter briefly outlines our Tiered approach to bot detection and builds on past research by adding a model that leverages a feature space extracted from 50,000+ entities collected with single seed snowball sampling. This model is developed for high accuracy but low volume applications. Our research shows that supervised machine learning models can leverage these rich structural, content, and temporal features associated with the target ego-network to increase model precision. Additionally, these network features offer an approach for modeling and detecting bot behavior that is difficult for bot *puppet-masters* to manipulate and evade.

Our evaluation of the bot-hunter suite of tools demonstrates that these models provide performance equivalent to or better than the state of the art. The Tier 1 model in particular is valuable to the community because it is accurate and can scale to large data (meaning researchers aren't required to sample their data). Additionally, because the Tier 1 model was designed to predict existing data, there isn't a requirement to return to the Twitter API to re-collect account data. This also means that it can be used to predict existing data sets that contain suspended or otherwise missing accounts.

Our analysis of Swedish political discourse on Twitter illustrates how bot-detection tools can support a typical open source intelligence workflow. The bot-hunter suite provides a way to enrich the data which can then be imported into other analysis tools for visualization and further analysis. Bot detection is not a "turn-key" solution and does require some work to set the right parameters, particularly the appropriate threshold level.

Chapter 4

Bot-Match

Social bots have emerged over the last decade, initially creating a nuisance while more recently used to intimidate journalists, sway electoral events, and aggravate existing social fissures. This social threat has spawned a bot detection algorithms race in which detection algorithms evolve in an attempt to keep up with increasingly sophisticated bot accounts. This cat and mouse cycle have illuminated the limitations of supervised machine learning algorithms, where researchers attempt to use yesterday's data to predict tomorrow's bots. This gap means that researchers, journalists, and analysts daily identify malicious bot accounts that are undetected by state of the art supervised bot detection algorithms. These analysts often desire to find similar bot accounts without labeling/training a new model, where similarity can be defined by content, network position, or both. To assist in filling this gap, we present the Bot-Match methodology in which we evaluate social media embeddings that enable a semi-supervised recursive nearest neighbors search to map an emerging social cyber security threat given one or more seed accounts.

4.1 Introduction

Today sophisticated state and non-state actors are using information systems in general and social media to change the beliefs and actions of target societies and cultures. These (dis)information campaigns, if left unchecked, gradually degrade the target society by eroding key institutions and values while widening existing fissures. This information "blitzkrieg" has led to the emerging discipline of social cybersecurity in which societies attempt to protect their culture and values from external manipulation while maintaining a free market for opinions and ideas. One of the key functions that computer science brings to the multi-disciplinary table of social cybersecurity is bot/cyborg detection and characterization.

Supervised and unsupervised machine learning models both provide important contributions to bot detection, but are not sufficient for social cybersecurity practitioners. Supervised models trained on specific labeled bot data provide an initial "triage" of social media streams, identifying likely areas of bot involvement and artificial manipulation of the online conversation. However, building supervised machine learning algorithms for every bot-detection scenario quickly grows untenable. Myriads of bot genres have evolved, including spam bots, intimidation bots, propaganda bots, social influence bots, cyborg accounts, and many others. Each of these bot genres have unique features and are curated and deployed in various ways depending on the target audience and culture. It is impossible to train a single model that generalizes to every genre, nor is it convenient to label and train models for every genre and then update these models on a frequent basis to keep up with bot evolution. Unsupervised learning, on the other hand, provides a way to find certain types of bots, such as the correlated bots found by the Debot algorithm [60]. These types of models are especially helpful in identifying labeled data for supervised models [32], but once again they are not sufficient. Often the most sophisticated and influential dis-information bots/cyborgs can fly "under the radar," undetected by either supervised or unsupervised models. Recently our team began to triage external manipulation of the Canadian political conversation in the run up to the Canadian 2019 national elections and found multiple influential accounts had emerged that were not being detected by production bot detection algorithms, but nonetheless were 1) divisive, 2) appeared to have foreign connections, and 3) appeared to have automated activity (i.e. were bots).

We find that social cybersecurity analysts in journalism, industry, academia and government, when faced with these sophisticated accounts, naturally ask the simple question "I wonder how many other accounts are similar to this one?" We developed *Bot-Match* to fill this gap, allowing analysts to rapidly find similar accounts in a flexible manner where the analyst can determine how they want to define similarity.

Similarity could be defined as similar network connections (either similar connections or similar network role), similar content, or a combination of both. With Bot-Match the analyst can choose to embed the conversation network, the conversation content, or both simultaneously, and then find similar accounts given a query. In this case the query is the seed node(s), and the algorithm returns the nearest neighbors given the predefined similarity measure. By recursively making this query, the analyst can rapidly build out a sophisticated information campaign that is undetected by other social cybersecurity tools. This approach is illustrated in Figure 4.1.

The contribution of this paper lies in its evaluation of graph and content embedding techniques in social media and its application of recursive nearest neighbors search to delineate latent groups of similar accounts. The paper also demonstrates the concept of querying social media data with an account as opposed to a keyword or hashtag. By using the rich features associated with a seed account, the Bot-Match algorithm provides fast and accurate information retrieval for social media analysts. This paper also validates this approach on social media associated with the 2018 US Midterm elections, and demonstrates the use of Bot-Match in a social cybersecurity workflow in support of the 2019 Canadian National Elections. From our research the Bot-Match approach is novel for social cybersecurity workflows and applications. While we are applying this in the specific context of malicious disinformation operations, the same methods can find and delineate any group of similar actors as long as the user has at least one seed account as an initial query. Beyond social media, this approach has applications in the broader problems of information retrieval, link prediction, and recommender systems (both collaborative and content filtering).

This chapter is organized as follows. It begins by describing past research in semantic and graph embedding, particularly focused on the models evaluated for inclusion in the Bot-Match methodology. We then conduct the formal evaluation of these models on two labeled data sets

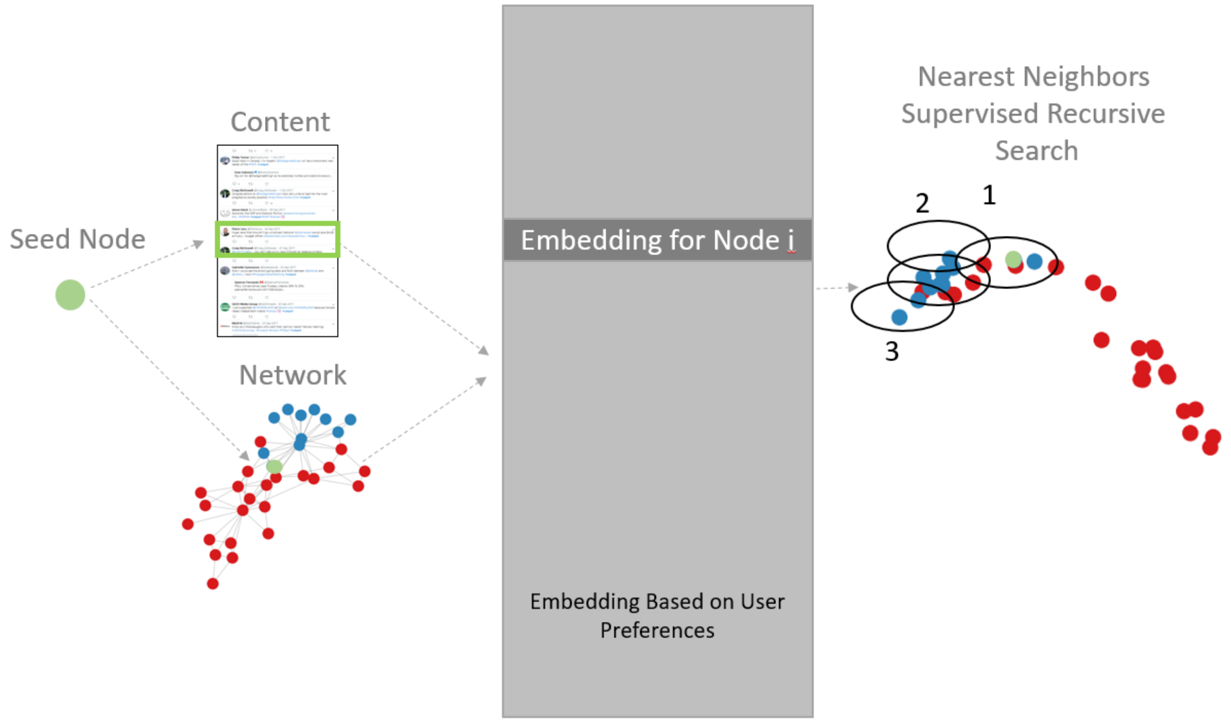

Seed Node
Content
Network
Embedding for Node i
Embedding Based on User Preferences
Nearest Neighbors Supervised Recursive Search
2
1
3

Figure 4.1: Framework to develop social media embeddings that enable a supervised recursive nearest neighbors search to map an emerging social cybersecurity threat given one or more seed accounts

associated with social cybersecurity, and use this evaluation to select models for Bot-Match. We then conduct a visual validation of the selected models using data associated with the 2018 US Midterm elections. Finally, we describe where Bot-Match fits in the social cybersecurity workflow and illustrate the use of the Bot-Match methodology in detecting disinformation actors in the 2019 Canadian national elections.

4.2 Review of Past Work and Evaluated Models

An embedding is a structure preserving map of one mathematical structure into another. The mathematical structure of X mapped into Y is defined as $f\ X \hookrightarrow Y$. In our case we intend to map semantic structure and graph structure into Euclidean space, both separately and then simultaneously. The embedding of semantic space has been an active research area since the 1950's [119] and graph embedding dates to at least the 1960's [235] with combinatorial approaches. In this section we will highlight past work and motivate our selection of evaluated models for semantic and graph embedding.

All of these models are transductive. The learned embedding allows us to find new data that is already represented within the network, but does not allow us to label new data from a different graph.

4.2.1 Semantic Embedding

Given the success of word embeddings [171], researchers have attempted to develop embeddings for phrases, sentences, and even documents. Early approaches simply averaged word embeddings for sentences and documents [192] which were expanded in sent2vec [188] using word and ngram embeddings while simultaneously training the composition and the embedding. Other approaches use a Recursive Neural Net (RNN) approach as demonstrated by the Skip-Thought model [141]. These were later trained over Natural Language Inference (NLI) to achieve improved results [233].

These approaches are often developed for a single language at a time, and are therefore limited on Social Media where most large conversations have multiple languages represented. Multilingual embeddings have been accomplished by learning jointly on parallel corpora [200] or by training independently and then mapping to a shared space with a bilingual dictionary [18]. Two competing models for universal encoding are 1) Google's Universal Sentence Encoder [58] and Facebook's Language-Agnostic SEntence Representations (LASER) toolkit [17]. The Google Universal Sentence Encoder (USE) encodes sentences and short paragraphs using two models, the Transformer model and the Deep Averaging Network (DAN) model. The Transformer model uses the encoding subgraph of the transformer architecture to create context aware embedding. The DAN model uses a feed forward deep neural network to average word and bigram representations. Facebook's Laser toolkit uses an encoder/decoder approach with Bidirectional Long Short-term Memory (BiLSTM) trained on 223 million sentences to create a universal encoding scheme for 93 languages. In our implementation Google USE was trained on cleaned and concatenated user content and Facebook Laser was trained at the individual tweet level and then tweet level embeddings were averaged to create a user/node embedding.

Prior to word and sentence embedding, researchers attempted to analyze topic groups with several methods, notably Latent Dirichlet Allocation [45] and Latent Semantic Analysis/Indexing [80]. Both Latent Direchlet Allocation (LDA) and Latent Semantic Analysis (LSA) are used to discover latent topics found in a corpus of documents and to reduce dimensionality. In the course of assigning documents to a fixed number of topics, both models are also reducing the dimensions of the corpus and inherently creating a document embedding. These approaches operate on a bag of words or tokens and are inherently multi-lingual (assuming appropriate language parsers). Latent Dirichlet Allocation (LDA) uses a probabilistic statistical model to map documents to topics while Latent Semantic Analysis/Indexing (LSA/LSI) uses singular value decomposition to reduce the dimensions thereby producing a set of topics or concepts. Both reduce the dimensionality of document representations and can therefore be used to embed text documents. We used term frequency-inverse document frequency (TF-IDF) for LSA, but term frequency for LDA since Blei describes how LDA was created to overcome some of the shortcomings of TF-IDF.

Finally, given that our end goal is measuring similarity, from the discipline of collaborative filtering we find another approach of measuring similarity and delineating neighbors without creating an embedding. Memory based collaborative filtering employ similarity measures to identify neighbors and thereby make recommendations for item-users data [204]. Using a similar approach we leverage Cosine and Jaccard similarity to measure similarity between users content based on a bag of words representation. Given a bag of words representation, cosine similarity measures the cosine of the angle between term frequency-inverse document frequency (TF-IDF) vector representations of the document. Jaccard similarity, on the other hand, compares the relative intersection of two documents. Note that cosine similarity is best if done on TF-IDF, whereas Jaccard similarity is performed on a bag of words representation. Cosine similarity will take into consideration frequencies, Jaccard similarity will only consider the presence or absence of words. This offers a baseline for comparison of more complex methods, and as we discover performs surprisingly well in our evaluation.

Table 4.1: Model Description by Type

Type	Subtype	Model	Data	Embed Dim
Content	Collaborative Filtering	Jaccard Similarity	Term Frequency	No embedding
		Cosine Similarity	Term Frequency	No embedding
	Topic Modeling	LDA	Term Frequency	N x 200
		LSI	TFIDF	N x 200
	Universal	Google USE	Combined Node Text	N x 512
		Facebook LASER	Text by Tweet	N x 1024
Network	Factorization	Graph Factorization	Adjacency Matrix	N x 32
		HOPE	Adjacency Matrix	N x 128
		BigGraph	Edge list - option for edge type	N x 1024
	Random Walk	node2vec	Edge list	N x 64
		Splitter	Edge list	N x 128
		role2vec	Edge list	N x 128
	Deep Learning	SDNE		N x 128
		GCN (no features)	Adj. Matrix & Bag of Words	N x 32
Network & Content	Deep Learning	GCN with Features	Adjacency Matrix	N x 32
	Factorization	BigGraph with initial	Edge list w/ LDA Embedding	N x 1024

4.2.2 Graph Embedding

While most graph based analysis is designed to operate on the original adjacency matrix or equivalent structure, recently methods have been devised to embed the graph in vector space. Vector space representations of graphs have applications in node classification, link prediction, clustering, and visualization [112]. Note that the notion of graph embedding in this paper is specifically focused on embedding nodes into vector space, not embedding the entire graph in vector space.

For the purposes of this research we've adopted the topology that Goyal and Ferrara introduced [112]. They divide graph embedding techniques into methods based on 1) Factorization, 2) Random Walk, and 3) Deep Learning. In our research we will test prominent models from each of these categories.

Factorization methods use various methods to factorize the adjacency matrix or other matrix representing the graph (Laplacian matrix, Katz similarity matrix, others). Eigenvalue decomposition can be used on matrices that are positive semi-definite, otherwise gradient descent methods are used. The primary factorization models we tested were the High-Order Proximity preserved

Embedding (HOPE) algorithm [187] and Facebook Biggraph [153], with Singular Value Decomposition of the Adjacency Matrix used as a baseline. HOPE preserves higher order proximity by minimizing $||S - Y_S Y_T||^2$ where S is a similarity matrix, instead of the adjacency matrix. In our case we used the Katz index to create the similarity matrix. Katz centrality is defined as

$$C_{katz} = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k \left(A^k\right)_{jk}$$

where A is the adjacency matrix and α is the attenuation factor (smaller than the absolute value of the largest eigenvalue of A).

In addition to using the HOPE algorithm, we also tested Facebook's Pytorch-Biggraph toolbox [153]. Pytorch-Biggraph can embed large graphs using several available factorization based models (TransE, RESCAL, DistMult and ComplEx). Pytorch-Biggraph overcomes complexity and memory constraints by partitioning the graph and then using multi-threaded and distributed execution with batched negative sampling. In testing Pytorch-Biggraph we wanted to determine what our loss of performance would be for scalability. Our results contain three settings for the BigGraph algorithm. The first setting is with a single edge type as is assumed in all other models. BigGraph allows the user to define different edge types, which was also tested and reported using our three edge types (*mention*, *retweet*, and *reply*). The third and final setting that we tested is BigGraph with an initial value, in our case a content embedding produced by LDA. This would initialize the training with knowledge learned from the content similarity, and as reported below always improved BigGraph performance (though in this setting BigGraph requires an initialization computation which may be computationally costly and which makes BigGraph no longer an end-to-end solution).

Multiple models leverage random walks to embed a graph. The model that we used is the *node2vec* model [116] which uses a biased random walk procedure to explore neighborhoods while preserving higher order proximity between nodes. We also tested the Splitter [89] random walk based algorithm that is tailored for social networks and embeds multiple persona-based representations of each user and then combines these to produce a single embedding for the user.

Within the random walk family of models we also implemented *role2vec* [9]. The *role2vec* algorithm leverages attributed walks which maps an input vector to a vertex type. The embedding structure for *role2vec* differs from all of our other methods in that it embeds the vertex type and not necessarily the vertex neighborhood. Our implementation uses the Weisfeiler-Lehman kernel [212] to extract vertex features.

Our primary Deep Learning model was the Structural Deep Network Embedding (SDNE) model [246] that uses deep autoencoders and decoders to preserve 1st and 2nd order network proximities. This is accomplished by optimizing both proximities simultaneously. The decoder is based on Laplacian Eigenmaps that penalizes similar vertices that are embedded far apart.

4.2.3 Content and Network Embedding

By its very nature social media data contains rich node features to include rich semantic features. Graph convolutional networks (GCN's) [140] have emerged as a way to embed a network simultaneously with the respective node features. While a variety of node feature representations exist in social media, we primarily used GCN's to simultaneously encode the network while considering node content features (the combined tweet text produced by each account). GCN scales better than SDNE by iteratively applying a convolution operator on the graph and aggregating the embedding of neighbors. The GCN defines a function of the form

$$f(X, A) = softmax\left(\hat{A}\ ReLU\left(\hat{A}\ XW^{(0)}\right) W^{(1)}\right)$$

where X represents node features and A represents the network adjacency matrix, $\hat{A} = D^{\frac{1}{2}}(A + I_N)D^{\frac{1}{2}}$.

4.2.4 Similarity Based Approaches to Bot Detection

Several past research efforts are somewhat related to our use of similarity measures to detect bots. One well-known and often cited unsupervised machine learning tool is the Debot algorithm [60] which uses warped correlation to find correlated accounts. These correlated accounts are bot accounts that post the same content at roughly the same time. Also notable is a model by Xiao et al. [256] that uses various features to classify entire clusters of new accounts on LinkedIn to detect batches of fake accounts. Magelinski et al. [162] demonstrates bot detection with graph classification by extracting a graph's latent local features and binning nodes together along 1-D cross sections of the feature space.

Finally, Ali Alhosseini et al. [13] uses GCN's with network, content, profile, and propagation features to train a supervised machine learning model to detect fake news given URL and cascade-wise detection. This study also explains and measures the decrease in bot detection performance as a model ages. Note that this research explicitly uses the GCN and geometric deep learning in a supervised manner, and while visualizing lower dimension representation of bot detection features, does not use these models in an unsupervised or semi-supervised manner as proposed by Bot-Match.

4.3 Data

4.3.1 Building Networks and Cleaning Text

Bot-Match assumes that the user has a filtered stream or conversation that she is trying to analyze for the presence of disinformation. This could be an online discussion around a topic (i.e. climate change), a political event (Canadian 2019 Elections), or a natural disaster (Hurricane Harvey). Note that in all cases the individual tweets are part of a larger connected online conversation and are not randomly selected from the social media environment (network analysis requires a network).

Below we list two data sets that we collected in order to evaluate Bot-Match. Bot-Match is designed to find similar bot accounts that have evaded a supervised machine learning initial approach. In order to test Bot-Match, we needed to find data sets where we can separately identify similar accounts, label them, and then test Bot-Match's ability to find them given a single seed node. The two data sets selected are discussed below:

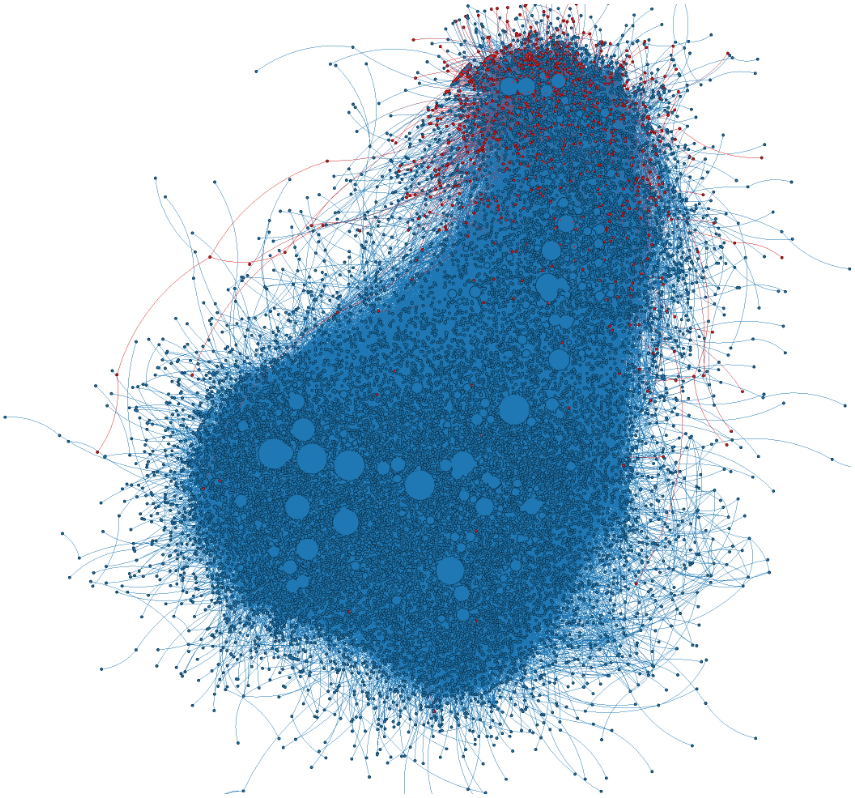


Figure 4.2: Conversational Network of Followers of a Journalist in Yemen. Red denotes random string accounts that were part of an intimidation campaign. Network includes 22,384 nodes and 189,379 edges.

4.3.2 Yemen Data

The first data set is the combined tweets produced by all followers of a freelance journalist in Yemen. Starting in the Fall of 2017, a determined and documented intimidation attack was launched against her Twitter Account [159]. The intimidation attack was characterized by a surge of strange accounts, many of them with strange and disturbing images or threatening messages. Many of these intimidation accounts were distinguished by 15 digit randomly generated alphanumeric strings for their screen name, such as **gG6RKc6QBqOLKyU** (not real screen name). We developed a logistic regression classifier to detect random strings based on features consisting of character n-grams and string entropy. Using this model we were able to achieve 94.25% accuracy in identifying random string accounts in Twitter, allowing us to automatically label 4,312 accounts characterized by a random string screen name and likely part of the coordinated intimidation campaign. While these accounts do not compose the entire intimidation attack, they nonetheless present an interesting social cybersecurity account that we can externally label and test Bot-Match performance in finding them given a query node. In our experiment, we will test and see if nearest neighbor searches of various embeddings would be able to find these random string intimidation accounts if we were not able to label them by their screen name. Throughout the rest of the chapter this data will be called the *Yemen* data. The conversational network of the

Yemen data is provided in Figure 4.2 with random string intimidation accounts colored red. Data description is provided in Table 4.2.

4.3.3 Internet Research Agency Data

The second data set consists of tweets produced by Russia's Internet Research Agency around the time of the 2016 US Elections and released by Twitter in October 2018 [237]. The St. Petersburg based Internet Research Agency (IRA) is a company that conducts focused online information operations on behalf of the Russian government and Russian businesses. The IRA represents one of the more experienced organizations involved in state-sponsored disinformation [36]. Twitter detected deliberate manipulation by the IRA, suspended the accounts and released the related data in an elections transparency effort (similar manipulation has been associated with Iran, Venezuela, China, and Spain).

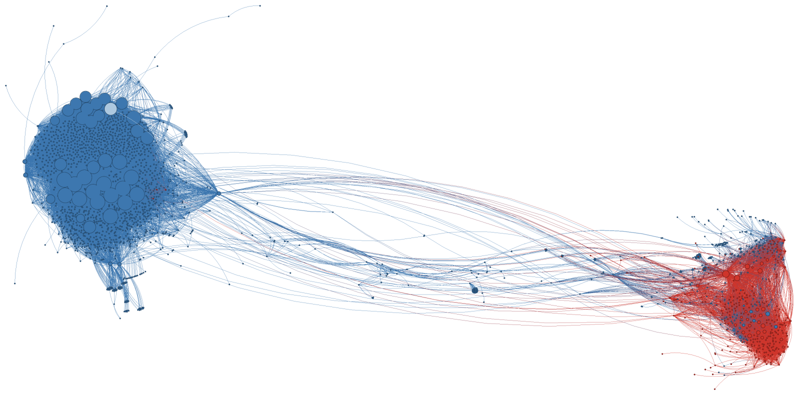


Figure 4.3: Conversational Network of Russian Internet Research Agency data released by Twitter. Red denotes accounts that targeted African American communities. Network includes 1,958 nodes and 35,931 edges.

The data demonstrates that the IRA specifically targeted African American online users to increase racial tensions in the United States [208]. For the purposes of testing Bot-Match we will label any account that shared relevant hashtags targeting African American populations as an account that is participating in this effort. In our case study we will test the performance of Bot-Match to detect these accounts after removing all hashtags. The conversational network of the IRA data (removing all nodes that didn't produce tweets in the released dataset) is provided in Figure 4.3. Data description is provided in Table 4.2.

4.3.4 Data Processing

Given curated, filtered, and related social media data, Bot-Match first builds the communication network edgelist by assigning a source and target for the directional communication. In the Twitter environment, this means creating directed links between accounts that *mention*, *retweet*, or *reply* to each other. Once the network is created, we also remove any nodes that are not in

Table 4.2: Data Summary

Data	Yemen Data	IRA Data
Tweets	4,535,117	9,041,308
Top Languages	en,ar,fr,es	ru,en,de,uk
Users/Nodes	35,763	3,667
Retweet Edges	108,382	31,398
Mention Edges	50,933	859
Reply Edges	35,857	4,122

the dataset (meaning we don't have content and node features for them). For example, a user may be mentioned in the data set but they never produced a tweet that ended up in the dataset. These users are therefore removed, as well as any isolates that remain. If keeping the users was required then collecting their content/timeline would be an additional data requirement.

For most of our models, *retweet*, *reply*, and *mention* edges are treated equally as directed communication links. Only Facebook's BigGraph algorithm will take into consideration the categories of links, with mixed results.

Unless otherwise noted (see comments on Facebook LASER), the text associated with each user (node) is a concatenation of all social media posts associated with that user. To clean the text we removed URLs, punctuation, reserved words, emojis and smileys. We removed hashtags and mentions with the IRA data but left them in for the Yemen data. They were removed from the IRA data since they were used to label the data and would artificially inflate content embedding models.

4.4 Model Evaluation and Validation

To evaluate embedding models for the Bot-Match methodology with the Yemen and IRA data, we created the respective *content*, *network*, and *network* + *content* embeddings for each data set, and then used these embedding to search for nearest neighbors of positively labeled accounts. For each data/embedding combination, we calculated the k nearest neighbors for $k \in \{10, 50, 100\}$ and then measured the precision of the response, defined as the proportion of positive responses. After conducting this query with each positively labeled account as the seed node, we averaged the precision across all queries to compute a metric for the given data/embedding combination.

Given that we calculated k nearest neighbors for $k \in \{10, 50, 100\}$, our primary metric was precision at $k = 10$ (p@10), precision at $k = 50$ (p@50), and precision at $k = 100$ (p@100). Given that these accounts aren't ranked, we cannot leverage any rank based metric, and precision at n is therefore appropriate. We can compare these percentages to the naive approach of random sampling. Random sampling with Yemen data gives precision of 0.12 and with IRA Data yields precision of 0.22. Any performance over these values indicates model value.

The results for the embedding test are provided in Table 4.3 and provide insights on both specific models and appropriate use cases for model *types*. The first observation is that all models provide significant value to the user when compared to naive baselines (Yemen: 0.12 and IRA:

0.22). For example, in the case of the Yemen data, if a user queries with a single random string intimidation account, on average 5 out of 10 returned accounts will be random string intimidation accounts (and the remainder may be other types of intimidation accounts used in this attack). This provides real and tangible value to analysts. The second observation is that all models perform better on IRA data than Yemen data. This is because the IRA data is smaller with a higher density of "similar" accounts and is more easily distinguished in both graph and semantic representation.

Table 4.3: Results. Note that the naive approach is to randomly select

		Yemen Data			IRA Data		
Type	Model	p@10	p@50	p@100	p@10	p@50	p@100
Content	Jaccard Similarity	0.421	0.387	0.380	**0.868**	**0.854**	**0.847**
	Cosine Similarity	0.371	0.303	0.287	0.835	0.808	0.790
	LDA	**0.457**	0.378	0.371	0.776	0.711	0.649
	LSA	0.424	0.359	0.340	0.796	0.780	0.754
	Google USE	0.426	0.373	0.364	0.716	0.681	0.659
	Facebook LASER	0.454	**0.406**	**0.393**	0.757	0.616	0.430
Network	Graph Factorization	0.391	0.258	0.239	0.625	0.549	0.486
	HOPE	0.342	0.283	0.275	0.715	0.539	0.409
	BigGraph (single edge type)	0.335	0.237	0.203	0.734	0.652	0.607
	BigGraph (multiple edge type)	0.299	0.212	0.182	0.722	0.652	0.606
	node2vec	0.365	**0.309**	**0.301**	0.734	0.658	0.617
	Splitter	0.258	0.172	0.149	0.663	0.605	0.506
	role2vec	0.326	0.248	0.231	**0.772**	**0.75**	**0.739**
	SDNE	**0.396**	0.303	0.271	0.701	0.606	0.545
	GCN (no features)	0.285	0.213	0.202	0.612	0.582	0.573
Network & Content	GCN with Features	**0.459**	**0.397**	**0.403**	0.685	0.632	0.618
	BigGraph with initial	0.356	0.250	0.213	**0.761**	**0.695**	**0.662**

Next we'll compare the embedding types. With the much more integrated conversation found in the Yemen data, we see the *content* models generally provide better precision across all values, while the more clustered IRA data has almost equal performance by both *content* and *network embedding*. The *network* + *content* embedding provides the best model for the Yemen data, and still outperforms most network models for the IRA data, albeit with a Biggraph as opposed to GCN.

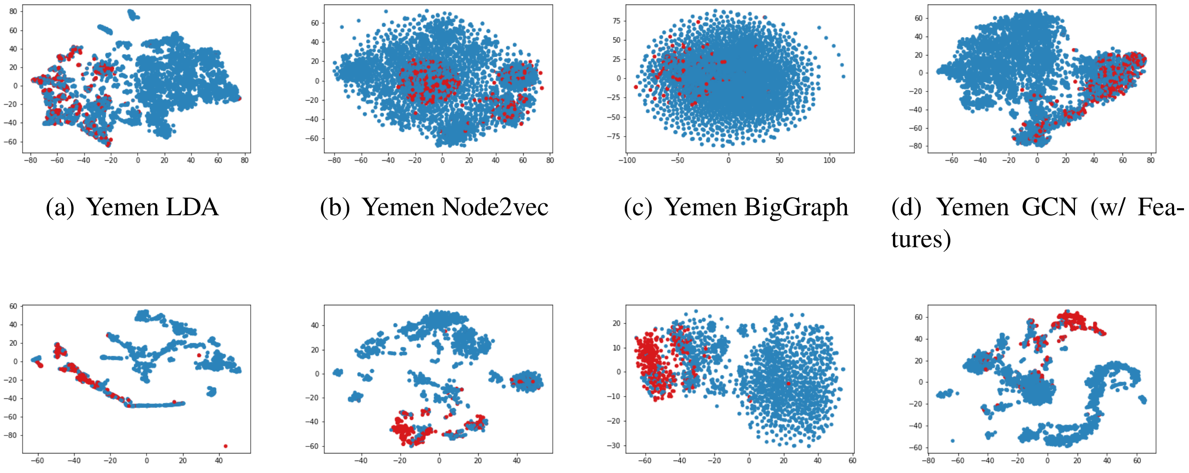
Focusing on the *content* algorithms, we see the classic models excel in similarity analysis. While these models will not necessarily create the contextual universal embedding that Google USE and Facebook LASER were designed for, they still excel at the basic task of document similarity and document retrieval. The LDA model and Jaccard similarity perform exceptionally well, are inherently multi-lingual based on a bag of words or bag of tokens (though care must be taken when choosing the size of the term frequency matrix in the presence of many languages).

Focusing on graph embedding, we see strong performance by random walk algorithms *node2vec* and *role2vec* across both datasets. Graph factorization and deep learning showed some success, though this fades at higher levels of n. The Pytorch Biggraph model scales much better than any other model, but did not perform as well as other models in our implementation. The GCN model without features and the Splitter model did not perform well in our evaluation.

Combining graph embedding with node features produced strong but mixed results. While GCNs produced the highest performance on the Yemen data, it produced mediocre performance on IRA data, with the reverse true for BigGraph (high results for IRA but less so for Yemen). This demonstrates that on this social data that the GCN is getting more traction on the content features as opposed to the BigGraph algorithms which is primarily focused on network features.

Given these results we selected Bot-Match algorithms for social media embedding in a social cybersecurity context. For our production Bot-Match algorithm, we offer LDA for content similarity, node2vec for network similarity, GCN with features for Network and Content embedding. We also use make Cosine Similarity and BigGraph implementations available for larger networks ($> 100K$ nodes).

In Figure 4.4 we provide t-Distributed Stochastic Neighbor Embedding (tSNE) visualization for the selected models for both the Yemen and the IRA data. These visualizations provide more insight into the model performances. We can visually see the higher precision of IRA data over Yemen data. We can also see various natural clusters emerging from the data, particularly the graph structure already visualized for the IRA data.



(a) Yemen LDA

(b) Yemen Node2vec

(c) Yemen BigGraph

(d) Yemen GCN (w/ Features)

(e) IRA LDA

(f) IRA Node2vec

(g) IRA BigGraph

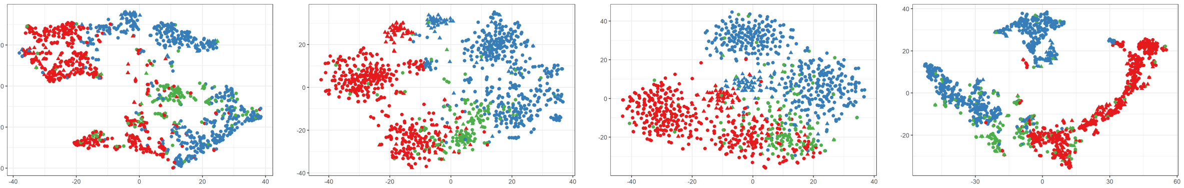
(h) IRA GCN (w/ Features)

Figure 4.4: t-Distributed Stochastic Neighbor Embedding (TSNE) 2-D Visualization of Embeddings

4.5 Visual Validation with 2018 US Midterm Social Media Data

Given the four models selected above, we wanted to conduct one additional visual validation of the Bot-Match methodology in general and these four models in particular. We had previously collected all Twitter content and connections associated with US Members of Congress or Congressional Candidates for the 2018 US Midterm elections. We decided to use this data to test our embeddings since it is easily labeled by both party (*Republican*, *Democrat*, *Other*) and by chamber (*House* or *Senate*). We wanted to test if the Bot-Match methodology and the four models selected above could leverage the social media connections (friend connections) and the social media content to capture the complex political environment of the US bicameral legislature in Euclidean vector space.

The data was prepared in the same way as the Yemen and IRA data, with the notable difference that the graph was constructed with *friend* links as opposed to *communication* links. Only members of Congress or Congressional candidates were retained as nodes in the graph. We then used our primary models as discussed above to create *content*, *network*, and *network* + *content* embeddings. Finally, we visualized these embeddings in two dimensions using t-Distributed Stochastic Neighbor Embedding (TSNE). The visualization of this is provided in Figure 4.5, where red indicates Republican, blue indicates Democrat, and green indicates another party. Circles indicate House politicians, and triangles represent Senate politicians.



(a) LDA (b) node2vec (c) BigGraph (d) GCN (w/ Features)

Figure 4.5: t-Distributed Stochastic Neighbor Embedding (TSNE) 2-D Visualization of US Congressional Members and Congressional Candidates for the 2018 US Midterm Elections. Red indicates Republican politicians, Blue indicates Democrat, and Green indicates Independent. Circle markers indicate House politicians/candidates, while triangles indicate Senate politicians/candidates.

From this visual validation, we see that all four models are able to capture similarity between politicians of specific parties, and within parties is generally able to separate members of the Senate from members of the House. All four embeddings are also able to identify specific factions with each of the parties. This visual validation gives us confidence that the selected embeddings are able to capture the rich graph and semantic features and map them to Euclidean space in such a way that a k-nearest neighbors search provides value in finding similar accounts.

4.6 Bot-Match Model in the Social Cybersecurity Workflow

Bot-Match is an important tool in the social cybersecurity workflow. While social cybersecurity workflows vary between teams and specific problem sets, a typical workflow is enumerated below. When an information campaign is initiated or expected, social cybersecurity analysts begin developing a data collection strategy in order to collect the core data associated with the information campaign or world event. Often this collection is either through key word filters or snowball sampling [111] of the network. Most large world events require iterative collection using both keyword filtering and snowball sampling. Once the data is collected, the team will begin with exploratory data analysis, which often consists of understanding the temporal and spatial density of the data as well as common hashtags and influential accounts. With exploratory analysis complete, analysts attempt to classify accounts and images. They often run bot/cyborg/troll detection as well as propaganda detection on the accounts. Bots are often used as force multipliers in an information campaign, and their presence can help outline the boundary of the campaign. Additionally, analysts may also attempt to classify actor type (media, politician, celebrity, etc). The classification stage can also include meme detection to extract all memes from the social media stream (memes are helpful in that they help clearly identify messaging and target audience). By the time the team finishes classifying accounts and images, they have usually whittled the data down to the data of interest, and now they begin a more intensive account characterization of the accounts in this smaller data set. Account characterization may be followed by campaign characterization, analysis of themes and narratives, and finally validation of campaign attribution, or identification of the perpetrator(s). These generic social cybersecurity steps are summarized and enumerated below:

1. Filter social media (key word filter or snowball sampling)
2. Exploratory data analysis (temporal/spatial distribution, common hashtags, influential accounts)
3. Classify accounts, images, etc
4. Characterize accounts, images, etc
5. Characterize campaign
6. Identify themes/messages/motives
7. Identify target audience
8. Attribution (identify perpetrator)

By the time that exploratory data analysis, account/image classification and account/image characterization are complete, the team has usually found a list of sophisticated accounts that are a core part of the information campaign. These core/interesting accounts become the input for Bot-Match, allowing the team to find similar accounts in the campaign, building out the information campaign in an iterative fashion.

In this way the Bot-Match tool and methodology is designed to be used in tandem with supervised machine learning models. Supervised models such as Botometer [74] and Bot-Hunter [33] can find large concentrations of bots, triage the network, and provide macro level bot statistics. However, many bots, often the most interesting and effective bots, remain undetected. This is caused by the fact that supervised models can be brittle and are biased by their training data toward specific bot types and genres [37, 261]. These accounts are often found through exploratory data analysis. Once found, Bot-Match allows the analyst to find many other similar accounts that have also likely avoided detection. The interesting accounts that Bot-Match returns to the analyst become new seed nodes, resulting in a recursive search pattern that allows an analyst to rapidly uncover sophisticated information operations in a matter of hours. This method of query is more effective than the key-word boolean search that is traditionally offered in social analytics tools. A query with all information (content and connections) is more useful than a query with a single relevant hashtag.

The embedding type (*content*, *network*, and *network* + *content*) is primarily selected based on user requirements. If an analyst wants to find accounts that post similar content as a seed account, regardless of where they are in the network, then they should leverage semantic similarity. If trying to find nodes that are proximate in the network, then network embedding is more appropriate. As a default, we found that embedding *network* + *content* with GCN (with Features) is a good default model if computationally feasible.

The most attractive attribute of Bot-Match is its ability to adapt to any problem or search requirement without labeling and training a new supervised machine learning model. All that is needed is a seed node and a target data set to search in. This provides tangible value to social cybersecurity analysts in particular and social media analysis in general.

In many ways the Bot-Match methodology provides a recommender system for social cybersecurity. Item-item recommendation systems (also known as collaborative filtering) recommend items based on similarity between the items, often measured by user ratings of those items. If you are interested in a hammer, then the recommendation system may recommend a hand saw based on item similarity. In our case, Bot-Match says that if you are interested in a certain account manipulating a target subculture, then you may also be interested in these additional k accounts that have similar connections and narratives. Selection of seed accounts could be done explicitly, or may be assumed through browsing and other search and exploratory actions.

4.7 Case Study

In this section we will illustrate the use of the Bot-Match algorithm and methodology in a social cybersecurity case study. In this case study, we will focus on analyzing information operations and specific suspicious accounts in the 2019 Canadian National elections. The 2019 Canadian National Elections were held on 21 October 2019. The formal campaign started on 11 September 2019, with a total campaign duration of 40 days.

Given the documented foreign influence in the 2016 US Elections [83, 176], the Canadian authorities took extra precautions to prevent similar tampering in their national election. The biggest policy they implemented was requiring all companies that have political advertising to set up a public facing registry with the specific ad and the name of the person who authorized that ad. Many companies (Reddit, Google, Microsoft, others) decided to ban political advertising altogether, while others (Facebook, Instagram, CBC.ca) began setting up registries [231]. This policy, while helpful in stopping manipulative paid content that the IRA leveraged in the 2016 US Election, does not stop manipulation by accounts that produce content that is not promoted through advertisement funding. This meant that bot, troll, cyborg, and sock-puppet accounts were still able to manipulate the conversation. Our goal was to find and analyze these malicious accounts.

To collect data associated with the 2019 Canadian National Election, our team used Twitter's streaming API to filter all tweets that contained hashtags associated with this event. We did this by starting with a few general hashtags associated with the election (i.e. #elxn43, #cdnpoli), and then weekly adding additional trending hashtags that we found in the data, finishing with 27 hashtags or tokens associated with the election. This collection produced 16.7 million tweets from 1.3 million unique accounts. The temporal distribution of the data with several major news events is provided in Figure 4.6.

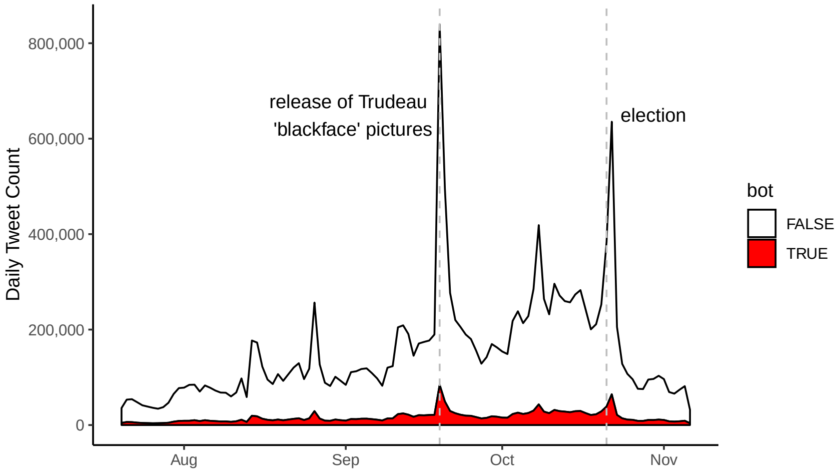


Figure 4.6: Temporal distribution of Canada 2019 National Election related tweets that were collected with the Twitter Streaming API. The density of accounts with "bot-like" attributes as predicted by the Bot-hunter tool [30] is shown in red.

Having collected the data, we built two embeddings for the data, one focused on the node embedding of the graph, and the other focused on content embedding for the content of the tweets. The scale of this data collection meant that some of the embedding techniques we explored were computationally difficult or impossible as implemented above. Given this, we used the Pytorch Biggraph model to embed the graph, and Latent Dirichlet Allocation (LDA) model to embed the content.

The Pytorch Biggraph model was used to embed the communication network created by directed links associated with the communication modes in Twitter (mention, retweet, reply). Latent Dirichlet Allocation (LDA) was used to embed the content. We found the Biggraph model was more computationally tractable (20 minutes vs. 2 days for LDA), but LDA provided more meaningful nearest neighbor relationships (the Biggraph embeddings provided too much noise in returned nearest neighbors).

Using the LDA model, we used Bot-Match methodology to find the 10 nearest neighbors of two sophisticated bot accounts that were manipulating Canadian political discussion on Twitter. One of the bot accounts was manipulating the political right (Canadian Conservative Party) and the other was manipulating the political left (Canadian Liberal Party). While this paper doesn't provide identifying information of the accounts, general descriptive information for both queries is illustrated and provided in Table 4.4. This table includes general information associated with the accounts (number or tweets, number of followers, etc), as well as bot prediction probabilities by two production supervised detection algorithms: Bot-hunter [30] and Botometer [74]. It also includes top hashtags by the query accounts and nearest neighbors to evaluate semantic correlation.

From Table 4.4 we see that all nearest neighbors of both queries are clearly associated with the stance and narrative of the query account as noted in the top hashtags. We also see that many of the accounts appear to have some automated activity (are a bot or cyborg account) as indicated by high volume and high retweet percentages. We also notice that many of these accounts were not detected by the state of the art production bot detection algorithms. The discrepancies between these two models, seen particularly in the second query, is likely due to very different bot genres used for training data.

As the analyst explores these accounts, additional accounts of interest may surface, creating new Bot-Match queries, which results in the recursive nearest neighbors search of accounts of interest. This recursive nearest neighbor search of graph and semantic embedding provides an important tool for social cybersecurity practitioners. I also want to highlight that Bot-Match is useful in finding "coordinating bots" that are sharing similar content with similar people. These "coordinating bots" are different than other bots, meaning that Bot-hunter models do not show strong performance when identifying these genre of bots.

Table 4.4: Descriptive Results of Bot-Match Queries of Sophisticated Bots Manipulating 2019 Canadian Political Parties.

Nearest Neighbors Query with Sophisticated Conservative Bot							
Screen Name	# Tweets	# Followers	# Friends	Bot-hunter	Botometer	Retweet %	Top Hashtags
Query	**39,396**	**11,157**	**4,616**	**0.690**	**0.148**	**0.636**	**TrudeauMustGo, cdnpoli, elxn43, DefundCBC, CPC**
Neighbor 1	41,212	3,635	4,992	0.457	0.691	0.678	TrudeauMustGo, Trudeau, Canada, cdnpoli, Canadians
Neighbor 2	7,886	935	1,715	0.378	0.148	0.592	TrudeauMustGo, cdnpoli, elxn43, LiberalsMustGo, SayNoToGlobalism
Neighbor 3	7,828	155	468	0.280	0.103	0.737	TrudeauMustGo, cdnpoli, elxn43, blackface, BREAKING
Neighbor 4	12,632	460	423	0.370	0.071	0.633	elxn43, TrudeauMustGo, cdnpoli, TrudeauWorstPM, TrudeauBlackface
Neighbor 5	5,939	385	256	0.398	0.129	0.546	TrudeauMustGo, cdnpoli, elxn43, LiberalsMustGo, ButtsMustGo
Neighbor 6	686	33	117	0.135	0.103	0.580	TrudeauMustGo, elxn43, cdnpoli, brownface, NotAsAdvertised
Neighbor 7	9,562	406	858	0.479	0.083	0.611	TrudeauMustGo, elxn43, cdnpoli, elxn2019, Scheer4PM
Neighbor 8	22,057	319	538	0.339	0.096	0.605	cdnpoli, TrudeauMustGo, LiberalsMustGo, elxn43, FakeNewsMedia
Neighbor 9	11	25	8	0.477	0.969	0.705	cdnpoli, elxn43, TrudeauMustGo, chooseforward
Nearest Neighbors Query with Sophisticated Liberal Bot							
Screen Name	# Tweets	# Followers	# Friends	Bot-hunter	Botometer	Retweet %	Top Hashtags
Query	**264,783**	**16,503**	**18,098**	**0.470**	**0.355**	**0.753**	**cdnpoli, elxn43, ChooseForward, topoli, onpoli**
Neighbor 1	11,791	394	160	0.335	0.071	0.816	cdnpoli, elxn43, ChooseForward, NeverScheer, TeamTrudeau
Neighbor 2	96,401	624	1,615	0.605	0.111	0.809	cdnpoli, elxn43, BREAKING, Scheer, Trudeau
Neighbor 3	10,432	206	1,528	0.740	0.083	0.842	cdnpoli, elxn43, ChooseForward, Elxn43, CDNpoli
Neighbor 4	35,557	735	379	0.560	0.071	0.870	cdnpoli, elxn43, ChooseForward, Trudeau, CPC
Neighbor 5	22,937	226	983	0.668	0.138	0.848	cdnpoli, elxn43, ChooseForward, ChooseForwardWithTrudeau, IStandWithTrudeau
Neighbor 6	27,377	1041	557	0.592	0.096	0.878	cdnpoli, elxn43, ChooseForward, TeamTrudeau, IStandWithTrudeau
Neighbor 7	5,026	438	1,540	0.558	0.103	0.843	cdnpoli, elxn43, ChooseForward, ScheerWasSoPoorThat, IStandWithTrudeau
Neighbor 8	3,445	334	1,026	0.564	0.066	0.768	cdnpoli, elxn43, ChooseForward, NeverScheer, YankeeDoodleAndy
Neighbor 9	37,845	257	387	0.435	0.222	0.927	cdnpoli, elxn43, ChooseForward, CPC, onpoli

As discussed above, we have already deployed prototype models of Bot-Match to monitor malicious disinformation in the Canadian elections. Having discovered sophisticated bot accounts manipulating both the political left and political right in Canada, we used Bot-Match to build out these campaigns and delineate the respective manipulative (dis)information operations.

4.8 Uses Beyond Social Cybersecurity

The concept of using an account and all associated features and connections as a query has many applications beyond social cybersecurity. These applications include retail, link prediction, intelligence, and information retrieval.

The retail business is one of the first adopters of recommender systems, and is arguably the most mature at deploying scalable collaborative filtering. These systems are inherently constrained by the user, and suffer from cold-start challenges. All product recommendations for the user are limited by the users own biases and ignorance, some of which they'd like to circumvent. Using the concept of an account query, an online retailer could allow a user to receive recommendations based on someone else's account (a celebrity, a friend, or someone else whose tastes they admire and wish to emulate). Social recommendations are some of the most powerful product recommendations, and allowing a person to receive recommendations as if they were someone else could be profitable for the retail industry. This approach would have a number of privacy hurdles to overcome, but if implemented correctly, allowing a customer to "Shop as if they were ...." could be the next big step in online retail.

Social media companies use recommendation systems and link prediction algorithms to make recommendations to a user based on their interests, content, and existing links. As with the retail business, some users may like to see recommendations as if they were someone else. For example, a young professional would like to see recommendations being presented to an established professional in their field who they follow and attempt to emulate. Once again, these raise significant but not insurmountable privacy concerns.

The final application area is in the area of intelligence. Many systems used for the intelligence community use simple boolean search patterns to search repositories of unstructured data. While key word searches may be required for initial exploration, as analysts find entities they are interested in, these entities and all content and connections associated with them could be used as a search query. This type of rich query could provide better search results for intelligence analysts.

4.9 Conclusions

This paper evaluates state of the art graphical and semantic embedding for social media data, and then leverages these embeddings for bot detection to enable social cybersecurity. Bot-Match is evaluated in two new social cybersecurity datasets, validated on a third dataset associated with US politics, and then demonstrated on a fourth dataset associated with the 2019 Canadian National Elections. Within the emerging discipline of social cybersecurity, the Bot-Match paradigm provides a novel way for analysts to find similar nefarious actors and recursively discover a complex disinformation operation without labeling and training a supervised machine learning model. Finally, while used within the social cybersecurity context, this approach has broad application to retrieval tasks that are characterized by network connections and semantic content. This includes document retrieval, recommendation systems, social recommendation, and other use cases.

Chapter 5

Meme Detection and Characterization

Combining humor with cultural relevance, Internet memes have become a ubiquitous artifact of the digital age. As Richard Dawkins described in his book The Selfish Gene, memes behave like cultural genes as they propagate and evolve through a complex process of ‘mutation’ and ‘inheritance’. On the Internet, these memes activate inherent biases in a culture or society, sometimes replacing logical approaches to persuasive argument. Despite their fair share of success on the Internet, their detection and evolution have remained understudied. In this research, we propose and evaluate Meme-Hunter, a multi-modal deep learning model to classify images on the Internet as memes vs non-memes, and compare this to uni-modal approaches. We then use image similarity, meme specific optical character recognition, and face detection to find and study families of memes shared on Twitter in the 2018 US Mid-term elections. By mapping meme mutation in an electoral process, this study confirms Richard Dawkins’ concept of meme evolution.

5.1 Introduction

Richard Dawkins first coined the word *meme* in his now famous book *The Selfish Gene* [77]. He developed the word meme by shortening the Greek word *mimeme* in an effort to create a “...noun that conveys the idea of a unit of cultural transmission, or a unit of imitation.” Dawkins indicated that memes function like genes for culture, and can undergo variation, selection, and retention. The meme is further defined as “an idea, behavior, style or usage that spreads from person to person within a culture” [43]. Examples of memes include shaking hands and singing “Happy Birthday”. As such, memes become building blocks of complex cultures [213].

Internet memes include any digital unit that transfers culture. This can be as simple as a phrase or hashtag, such as the *Diasoi* meme in China [227] or the #MeToo movement in America. The Internet provides an environment for digital memes to quickly move from person to person, often mutating in the process as initially envisioned by Dawkins. In 1982 the first emoticon (:-)) was used on Carnegie Mellon University’s online bulletin board in order to flag humor [76]. As a merger of humor, text, and a symbol, the emoticon became one of the first types of Internet memes.

Figure 5.1: Memes used in conjunction with the US 2018 Midterm Elections.

While Internet memes can exist as words, emoticons, videos, or gifs, a common form is an image with superimposed text that conveys some type of merged message. In the earlier days of the Internet, images with superimposed text began to propagate via *Usenet*, email, and message boards. By the early 2000's researchers began to study these specific visual artifacts that were proliferating. Social networks soon emerged, allowing these memes to go viral.

Given the power of memes to appeal to cultures and sub-cultures, various political actors increasingly use them to communicate political messaging and change the beliefs and actions of the fabric of a society. Canning even goes so far as to claim that memes have replaced nuanced political debate [53]. Memes become a simple and effective way to package a message for a target culture. Memes are used for politics, magnify echo chambers, and attack minority groups [190]. This has jumped into the public discourse with various articles, including the *New York Times* article "The mainstreaming of political memes online" [49]. The increasing use of Internet memes for "information operations" has led to our effort to detect and characterize memes that inhabit and propagate within given world events and the conversations that surround them.

Few research efforts have attempted to capture a comprehensive dataset of political memes and the network they travel through in a political election event, and then document how the memes evolve, propagate, and impact the network. Our work will develop a deep learning method to detect memes in social media streams and leverage graph learning to cluster these images into meme "families". We will then apply these methods to Twitter data streams associated with the 2018 US Mid-term elections and the 2018 Swedish National Elections. In addition to contributing a theoretical framework for classifying and clustering meme images, our research indicates that memes are shared less but move to more places on the Internet when compared to non-meme images. Memes therefore spread through different mechanisms than other "viral" content.

This chapter is organized as follows. In the Section Related Work, we describe the history of the Internet memes, prior work exploring data analysis approaches to study memes, and deep neural networks that have been used on similar problems. Then in Section 5.3, we propose Meme-Hunter, a deep learning model to find images on the Internet and classify them as meme vs. non-memes. We then use the models to study the usage of memes in two elections in Section 5.4. Finally, we conclude the findings of this research and suggest directions for future work.

5.2 Related Work

5.2.1 History of Internet Memes

The study of memes has existed ever since Richard Dawkins introduced the concept in his book 'The Selfish Gene' in the 1970's [75]. Many researchers have attempted to study the relationship between memes and culture. The advancement in Internet technologies and the world-wide-web (www) gave meme researchers a laboratory with which to study the spread and mutation of memes. This led to several books on memes, the most influential and controversial being Blackmore's *The Meme Machine* [42, 214].

Linor Shifman has conducted extensive research of digital memes from the perspective of journalism and communication. In 2013 Shifman deviated slightly from Dawson's original definition and defines the Internet meme as artifacts that "(a) share common characteristics of content, form, and/or stance; (b) are created with awareness of each other; and (c) are circulated, imitated, and transformed via the Internet by multiple users" [215, 216]. She also differentiates viral content from memetic content. She claims that *viral content* "is defined here as a clip that spreads to the masses via digital word-of-mouth mechanisms *without significant change*." In contrast, *memetic content* is "...a popular clip that *lures extensive creative user engagement* in the form of parody, pastiche, mash-ups or other derivative work."

In 2012 Davidson observes and discusses the fact that Internet memes typically lack attribution [76]. Unlike many other creative works, authors of Internet memes typically don't attach their name to the memes they create. They remove any traces of attribution from the file and its metadata, and usually introduce memes on sites that offer anonymity (4chan, Reddit, etc.), where they gain popularity before hopping over to mainstream media (Facebook, Twitter, etc.) [23]. Several theories exist that explain this behavior, but Davidson seems to offer the most logical in that anonymity enables a type of freedom. This freedom allows authors to create and distribute questionable material without concern for retribution from authorities. It is this lack of certain attribution that encourages malicious and divisive political actors to resort to memes for information operations.

The far-reaching impact of a meme's evolution combined with the often inherent anonymity make memes attractive to various political and propaganda campaigns. The evolutionary nature of memes assists them in 'hopping' platforms to move to additional Internet and social media spaces. The natural anonymity of memes allows various actors to make it appear that the distribution of the messages is part of a grass roots movement. Donovan and Friedberg discuss how images can be used to as “evidence collages” in a “source hacking” operation [84], thereby providing seemingly legitimate evidence of a false event or biased conclusion. It is these aspects of political and propaganda memes that we want to apply our research.

5.2.2 Meme Detection

Deep neural networks (DNN) have shown great success in many fields [126]. Researchers have used DNNs for various vision tasks like the Imagenet Challenge [82, 143] and fashion recommendation [220]. DNN’s have also been used for various natural language processing (NLP) tasks like Part of Speech (POS) tagging and named entity recognition [67]. Ironically, deep learning has more often been used to automatically generate Internet memes as opposed to find them. In 2013 Wang et al. [249] used copula methods to jointly model text and vision features with popular votes. In 2018 Peirson et al. [190] leveraged deep learning to generate memes in a model they titled “Dank Learning”.

Xie et al. [257] used YouTube to find short video segments that are frequently reposted which they call video memes. The authors then created a graph of people and content to model interactions. Unlike video memes, exploring image memes is more challenging as this requires first classifying an image as meme or not-meme.

The closest research related to our detection effort is the Memesequencer model developed by Dubey et al. [85]. This research separates the meme image template (underlying image) from the additional text and image manipulation. After separating the meme template it creates a meme embedding by concatenating image features and text features using deep learning, with the best model concatenating ResNet18 with SkipThought text features. Having created an embedding, the authors construct the evolutionary tree using a phylogenetic tree. This research is limited to memes that have identifiable templates found on sites like Memegenerator or Quickmeme. When used to extract memes for social cybersecurity practitioners, the Memegenerator provides high precision but low recall (see below). Our intent with Meme-Hunter is to increase recall.

5.2.3 Meme Evolution

The digital footprint that Internet memes leave allows researchers to study the propagation of memes through (and across) networks. Coscia looked at meme propagation and measurements of success in 2013 [69]. Bauckhage et al. [23] explored the temporal models of fads by looking at Internet memes, approximating interest in a given meme by using Google Trends. Leskovec et al. [154] used memes and phrases extracted from news and blogs to track and study the dynamics of the news cycle. This work was able to map the evolution of text-based memes in the news cycle and blogosphere. Ferrara et al. [94] focused on clustering text-based memes.

The closest research to our study of meme evolution is the study by Zannettou et al. [264] that clusters image streams based on pHash and identifies memetic clusters using meme annotation from sites such as “Know Your Meme”. They apply this process to multiple sources (Twitter, Reddit, 4chan, Gab) and then use Hawkes process to measure which ecosystem has greater influence. While focused on meme evolution and influence, this paper does not specifically develop a detection model that generalizes easily beyond the Know Your Meme annotations, once again rendering low recall in detection applications. Additionally, this paper clusters only based on the image (via pHash) and does not consider the multi-modal nature of memes when measuring similarity.

5.2.4 Meme Optical Character Recognition

The classification process requires learning from a composition of image and text characteristics. Extracting text in memes requires Optical Character Recognition (OCR). OCR on memes can be challenging since most OCR algorithms are trained to recognize black font on white background, where many memes are white font on dark background. For social media image OCR, the state of the art is arguably the Facebook Rosetta system, a deep learning model that conducts OCR while taking into consideration the background as well [48]. This is being deployed on Facebook's platform in order to censor images for extremist messages, allowing Facebook to comply with increased regulation, particularly in the European Union. Facebook Rosetta output is standard OCR output (text), and it is not intended to classify memes vs. not-memes. It is also not open sourced or available for researchers (at the time of this writing).

Our research combines some of the efforts of Zannettou et al. [264] with that of Dubey et al. [85]. In doing so, we go beyond both papers by creating a generalizable multi-modal meme detection model that is not constrained by annotated entries on a site like Know Your Meme. Additionally, we develop the evolutionary graph with a radius nearest neighbors approach and apply this specifically within the online debate around a large election event (2018 Mid-term elections). This provides the research community with a generalizable multi-modal meme detection model, a new way to build an evolutionary tree, a meme OCR pipeline, and insights into meme impact and propagation within political conversations. Additionally this model provides approximately 8 times increase in recall over the template-based methods that Dubey and Zannettou propose. This increase in detection recall is especially important for social cybersecurity practitioners.

5.3 Classifying Images as Memes

Most images shared on platforms like Twitter are not memes (see Table 5.3 for stats). Therefore, to explore the usage of memes, it is essential to first classify if an image is a meme or not. While visual Internet memes come in a wide variety of formats, we restricted our classification to two types that are commonly found. These two types are found in Figure 5.4 and can be described as:

1. A picture with superimposed white text in impact font. Impact font was developed in the 1960's by Geoff Lee and is the font of choice for text over image [87]. This is illustrated in Figure 5.4(a)
2. Text placed in a white space over a picture, as is shown in Figure 5.4(b)

While this seems restrictive, we will show later that, even with this constraint, our approach finds $8\times$ more memes (i.e. $8\times$ higher recall) than template-based methods.

Given enough meme vs non-meme data, it could be possible for a neural-network model to learn to extract text (using OCR), extract faces and other meme characteristics to classify a meme. However, in a limited data setting like ours, this approach is likely to fail as OCR itself is research domain in itself. Consequently, we propose to first extract text and face encodings and use them as supplementary input features. Then to predict an image to be meme/non-meme, we explore deep learning based multi-modal (multiple features) models that use extracted features in addition to the raw images.

Next, we describe our models, our data collection effort to get meme and non-memes data to train the models, the process of training the models, and the classification performance.

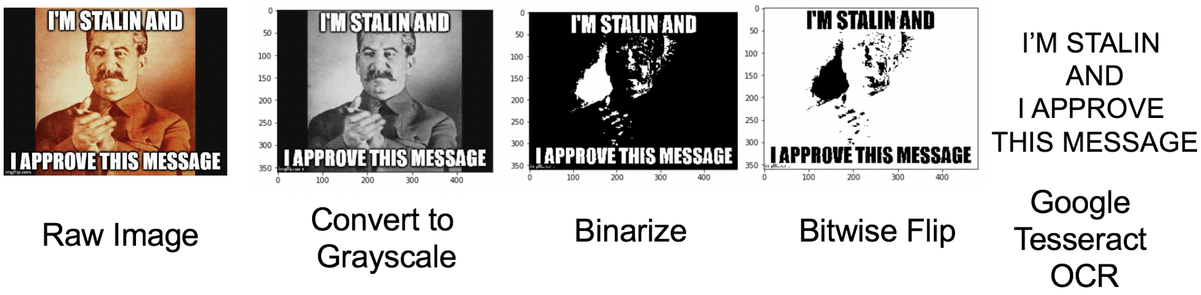


Figure 5.2: OCR Pipeline for Meme Images.

5.3.1 Memes Classification Models

As mentioned earlier, we first extract text and face encodings, so here we explain the process of extracting text and face encodings from images.

Text Extraction For Optical Character Recognition (OCR) we combined meme specific image preprocessing with an open source OCR tool. When images contained white font over dark background, we preprocessed the images by 1) converting the image to grayscale, 2) binarizing the image, and 3) inverting every bit in the binary image array. These image preprocessing steps are illustrated in Figure 5.2. OCR on preprocessed images was accomplished with Google Tesseract [219]. If images already had black text on white background, no preprocessing was applied. Our experiments indicated that preprocessing significantly improved Tesseract's OCR on meme images. Baseline Tesseract required an average of 49.8 ± 13.8 character edits (or levenstein distance) with only 2% readability. Preprocessing reduced this to an average of 17.5 ± 4.8 character edits with 72% of strings remaining readable.

Human Face Encoding As faces are an important element of memes, we extract facial features using the open source face detection software package called *face_recognition*, created by Adam Geitgey and made available at [105]. The library returns a face encoding vector for each face found in the image. We use these vectors as the input to our classification models.

We tried four different groups of classifiers: 1) unimodal classification using only text 2) unimodal classification using only machine vision 3) multimodal classification using text and vision, and 4) multimodal classification using text, vision, and face encoding.

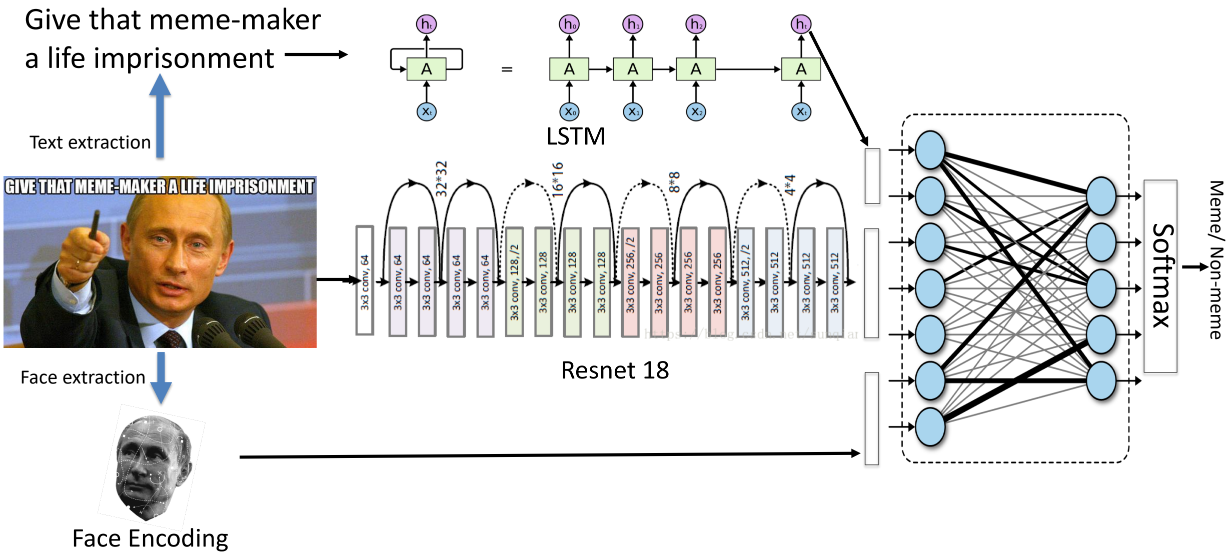


Figure 5.3: Joint Model for meme classification.

LSTM-based text classifier In this unimodal model, we use only the extracted text from images as the input for meme classification. Long Short Term Memory (LSTM) [127] networks are very popular for text classification. An LSTM takes word embedding and a hidden vector as the input and outputs a new hidden vector. At the end of the text (input), a fully-connected layer followed by a softmax layer is used to predict the label of the text. We used Glove vectors [192] as the input word embeddings. In our results, we provide several other text only models for comparison, including Naïve Bayes, Support Vector Machines, and Logistic Regression.

CNN-based image classifier Given that our work focuses on image-based meme detection, and Convolutional Neural Networks (CNNs) are the most popular models for visual learning, it is natural for us to consider a CNN based model. For this work, we tried a number of pre-trained models including VGG18 [217], ResNet18 [120] and ImagenetV3 [228]. For classification, we removed the last fully connected layer of the pre-trained network, added a new fully connected network followed by a sigmoid layer. We also explored freezing all layers, freezing some of the layers, and not freezing any of the layers in the training process. In the end, allowing to update the weights on all layers provided the best results. We also include results that extract descriptors with scale-invariant feature transform (SIFT) and Bag-of-Visual-Words (BOVW) feature representation and support vector machine classification. The SIFT-BOVW model is provided to demonstrate DNN improvement over pre-DNN models.

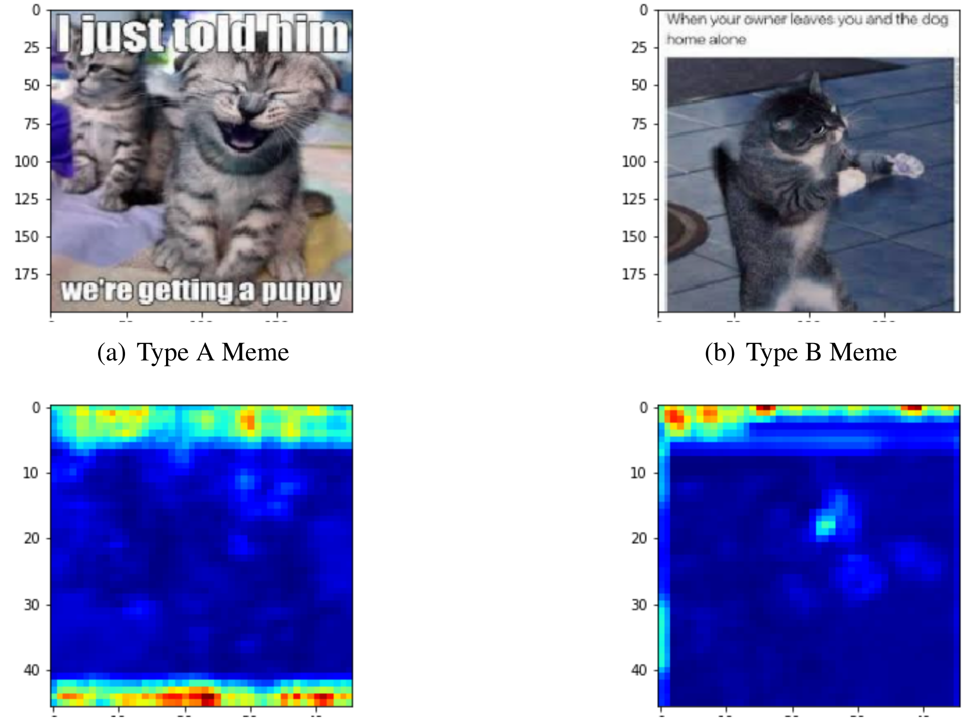
Joint DNN model The joint DNN model approach starts by combining just the LSTM (discussed above) and CNN (discussed above), and then combines the LSTM/CNN with face encoding features as a single model. The model's architecture is shown in Fig. 5.3. As shown in the figure, the output of the LSTM, the CNN and face encodings are concatenated as a single vector. The concatenated vector is then used as the input to a dense fully connected layer followed by a sigmoid activation. All parts of model are trained jointly.

In the last connected layer we use a sigmoid (or logistic) function to generate a probability of the image being a meme. The sigmoid function is defined as

$$\phi(z) = \frac{1}{1 - exp(z)}$$

5.3.2 Data

To label Internet meme images for supervised learning, we searched meme images on Reddit, Twitter, Tumblr, Google Image Search, Flickr, and Instagram. Collecting images from these platforms, we were able to find 25,109 meme images. The meme data contained varied meme categories, including sports, politics, celebrities, and animals. While the dominant language is English, other languages include French, Spanish, German, Russian, Japanese, Arabic, and Chinese. The *non-meme* images were collected at random from Twitter and Google Image search.



(a) Type A Meme

(b) Type B Meme

(c) Saliency in Type A Meme

(d) Saliency in Type B Meme

Figure 5.4: Two types of memes used for meme classification with their respective saliency maps. Saliency maps are computed by averaging pooled gradients across channels.

In the training data we filtered out non-meme images that didn't contain either text or a background photo. This was done so that the algorithm would learn the unique attributes of meme images as opposed to just learning to identify the presence or absence of text. In order to filter for text, we needed to conduct text detection but not necessarily text recognition. We found that the Efficient and Accurate Scene Text (EAST) detection model [268] performed better at detecting text than the Tesseract based OCR pipeline discussed earlier. Note that the EAST model detects the location of text in an image but does not recognize or extract the text. We used the EAST algorithm to filter out any images that didn't have at least one text bounding box. Having removed images that don't contain text, we discovered that we also needed to remove images that don't contain a photograph. This decision was made after finding many black and white document images, particularly in political conversations. To remove document images, we developed a heuristic that measured the mean Red Green Blue (RGB) score for the image, and removed it if the mean score was greater than 220. This proved to be fast and easily removed document images without removing memes of interest. This filter was applied in both the training process as well as the production algorithm.

$$\textit{Image is document if} \quad \frac{Red + Green + Blue}{3} > 220$$

Table 5.1: Classification Dataset Statistics.

Total Images	Memes	Non-memes
50,209	25,109	25,100

We summarize the final model training dataset in Table 5.1. The 50,209 images were mixed with equal portions of *meme* and *not-meme* images. The data was then randomly split into *training* data (80%), *validation* data (10%), and held out *test* data (10%).

Collecting images from social media streams often includes some amount of abusive language and adult content images. Practitioners using our methods who want to minimize the impact of this sensitive content should have an appropriate filter. In our case we used Yahoo's Open Source "Not Safe For Work" (NSFW) filter (`https://github.com/yahoo/open_nsfw`).

5.3.3 Experiments and Results

For the meme classification task, we define the overall objective function using cross-entropy loss, as can be seen in Equation 5.1, where $i \in n$ samples, $j \in \{meme, non\text{-}meme\}$ classes, y is the (one-hot) true label, p is the probability output for each label.

$$L(y, p) = -\frac{1}{n} \sum_{i,j} y_{ij} \log(p_{ij}) \tag{5.1}$$

Our primary metric of interest is the F1 score, defined as the harmonic mean of precision and recall. We used this as our primary metric since it balances the often competing priority of precision vs. recall. In our results we also provide accuracy, precision, and recall for interpretability.

All models are built using the Keras library[1] with a Tensorflow backend [2]. As described earlier, the models use text, face-encoding, and image features as the input and a sigmoid layer

[1] https://keras.io/

[2] https://www.tensorflow.org/

for the class label prediction. The models are trained using stochastic gradient descent with a cross-entropy loss function as seen in Equation 5.1. The learning rate was used as a hyper-parameter and varied from 10^{-3} to 10^{-1}. The LSTM hidden layer size was varied from 16 to 256. We found that a hidden layer size of 50 and a learning rate of 10^{-3} worked well. These hyper-parameters were then fixed during the training and testing process.

Table 5.2: Classification Results.

Type	Model	Accuracy	F1	Precision	Recall
Text	Logistic Regression	0.724	0.719	0.735	0.703
	Naïve Bayes	0.681	0.607	0.793	0.492
	SVM	0.721	0.714	0.736	0.693
	LSTM	0.799	0.805	0.786	0.824
Vision	SIFT-BOVW	0.798	0.788	0.828	0.752
	Baseline CNN	0.939	0.938	0.946	0.930
	VGG18	0.915	0.916	0.909	0.923
	ResNet18	0.926	0.927	0.907	0.948
	Inception-V3	0.958	0.958	0.952	**0.964**
Multi-modal	Vision + Text	0.954	0.954	0.943	0.965
	Vision + Text Length	0.952	0.951	0.947	0.956
	Vision + Text + Face	**0.961**	**0.961**	**0.959**	0.963

We compare the performance of the models in Table 5.2 and show the training plots in Figure 5.5. We train the models for only 10 epochs since the performance plateaus after that. As we can observe from the plots, most of the learning is done in the first epoch and validation accuracy is high thereafter. From these results we see that the LSTM model is significantly better than other text models. Within the Vision models, we see that all DNN models show significant improvement over the SIFT-BOVW model, with the Inception-V3 very deep model providing the best performance across all metrics. We do see that the multi-modal models provide slight improvement over unimodal vision models. Model saliency maps [218] are provided in Figures 5.4(c) and 5.4(d). Saliency maps show the salient pixels that are important for a given class and are computed by averaging pooled gradients across channels. From these saliency maps we see that we are indeed learning to identify images where the text is positionally located in pixel locations that are indicative of meme images. Overall we can summarize results by claiming that unimodal machine vision models provide solid performance in meme detection and can be enhanced (at a computational cost) with multi-modal text-based features.

5.4 Evaluating Memes in Election Events

5.4.1 Finding Memes

We used the DNN model to classify images used in the 2018 US Midterm Elections and the 2018 Swedish National Elections. We will focus on the 2018 US Midterm election data because

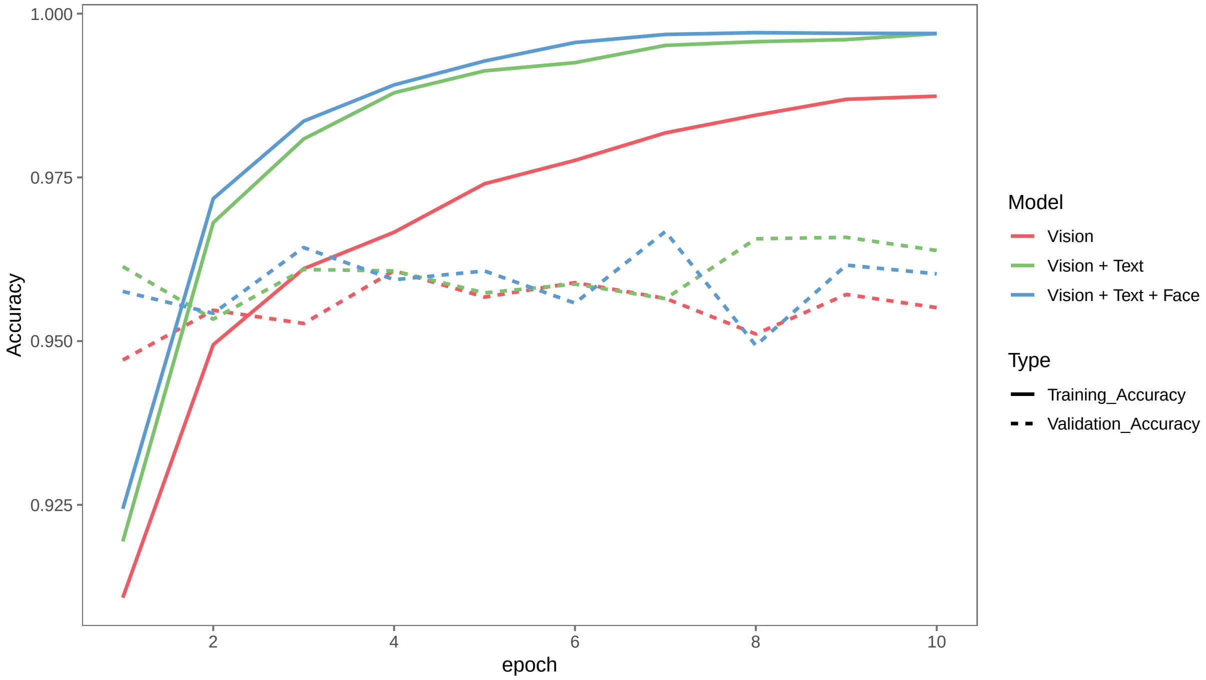


Figure 5.5: Comparing training and test performance of different models.

it provides the largest meme collection, but the 2018 Swedish election data is provided in Table 5.3 for comparison purposes. For the US Midterm elections, we collected all tweets that mentioned a member of congress or congressional candidate. For the Swedish elections, we collected tweets containing hashtags associated with anti-immigrant and nationalistic movements (#svpol, #Val2018, #feministisktInitiativ, #migpol, #valet2018, #SD2018, #AfS2018, and #MEDval18). Note that the Swedish election data does not cover the full spectrum of politics in Sweden, but the US Midterm election data does cover the full spectrum of politics in the United States. We downloaded all images from both data sets in February 2019. As indicated below, approximately 9% of the images weren't available (the account or tweet was suspended by Twitter or removed by the account owner). The statistics for both data sets are provided in Table 5.3.

We conducted binary classification with our trained DNN model on all images extracted from both data streams. A collage of examples that we classified as memes in the US mid-term elections is provided in Figure 5.1.

5.4.2 Mapping Meme Evolution in Political Conversations

Given the rich vision/text data that we had, we wanted to map the evolution of visual memes using similarity clustering. By clustering these images, we can not only identify the families but also the connections between the families of memes. We explored several proven methods for measuring image similarity, to include Color Histograms [184], Scale-Invariant Feature Transform (SIFT) [160], Perceptual Hashing (pHash) [59], and a method similar to the Deep Ranking [248]. Similar methods have been used with K-nearest neighbors for image annotation [225] and with mapReduce by Google for clustering billions of images [158]. Our initial experiments

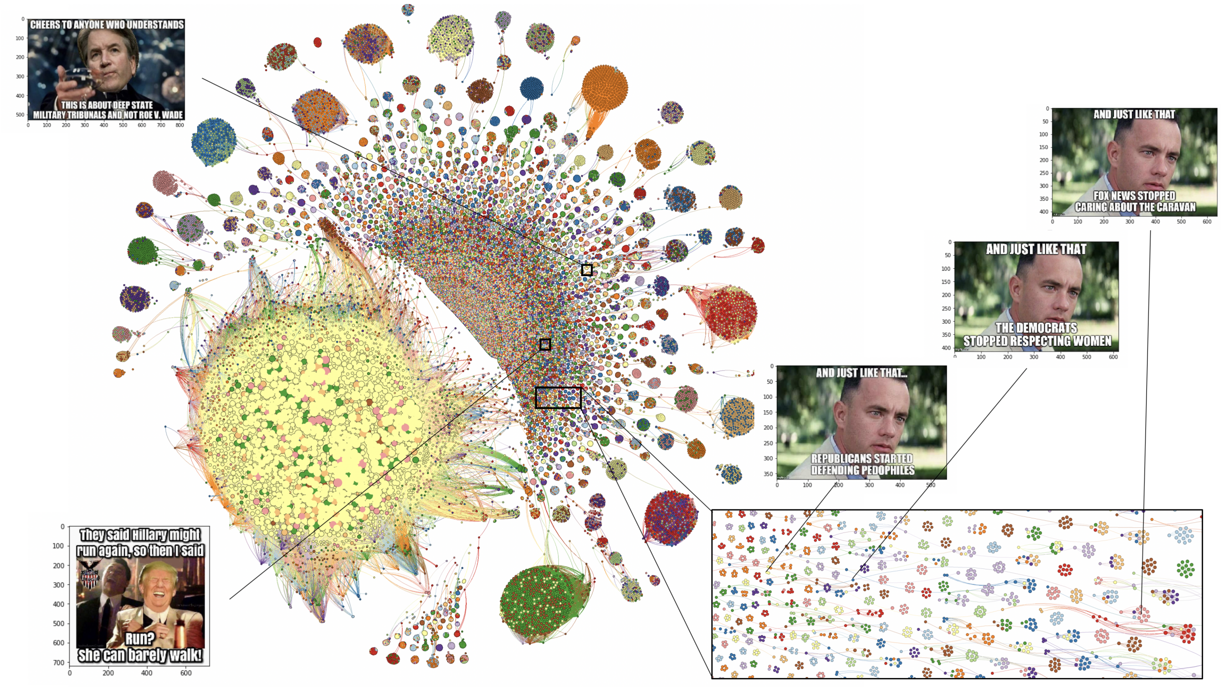

CHEERS TO ANYONE WHO UNDERSTANDS
THIS IS ABOUT DEEP STATE MILITARY TRIBUNALS AND NOT ROE V. WADE
AND JUST LIKE THAT
FOX NEWS STOPPED CARING ABOUT THE CARAVAN
AND JUST LIKE THAT
THE DEMOCRATS STOPPED RESPECTING WOMEN
AND JUST LIKE THAT...
REPUBLICANS STARTED DEFENDING PEDOPHILES
They said Hillary might run again, so then I said
Run? She can barely walk!

Figure 5.6: Graph Learning with Fixed Radius Nearest Neighbors showing *families* of memes in the US 2018 mid-term elections (89K nodes and 1.87M links). Network visualization is done with Graphistry (`https://www.graphistry.com/`).

reveal that the deep ranking method (using features extracted from the last layer before softmax and evaluated with euclidean distance) performs well. To identify the families of memes, we finally used graph learning with fixed radius nearest neighbors algorithm [29]. Fixed radius nearest neighbors finds the neighbors within a given radius of a point or points. We chose fixed-radius method over the K-nearest neighbors method since the size of our meme families vary widely. This technique also allows us to quickly query similar images based on a fixed distance radius.

Given a meme, we used ‘brute-force’ based radius neighbour algorithm to find the mutations of the meme. We attempted to use the ball tree algorithm [186], which partitions meme features into a nested set of fixed dimensional hyper-spheres (balls) such that each hyper-sphere contains a set of memes based on its distance from the ball’s center. Although the ball-tree was designed for high dimensionality, we found that this is still computationally expensive with more than 120 features. With 25,088 features, we found that the ball-tree algorithm was not practical, and resorted to the brute force algorithm. Once we have the neighbours of a meme, we can use time of the posting associated with the meme to generate a directed graph of meme mutations. We recurse the whole process over the neighbours to get the next set of neighbours and add them to the graph. We stop the recursion after a fixed set of steps or if the max size of the graph is attained. The algorithm is summarized below (Algorithm. 1). The map of all nodes and links for the 2018 US Midterm elections is provided in Figure 5.6. In this we clearly see the clusters of similar images (or “families”), as well as some of the connections between them.

Having mapped the individual “families” of memes, we used this similarity clustering and the date-time information from the Tweet metadata to map the chronological evolution of specific

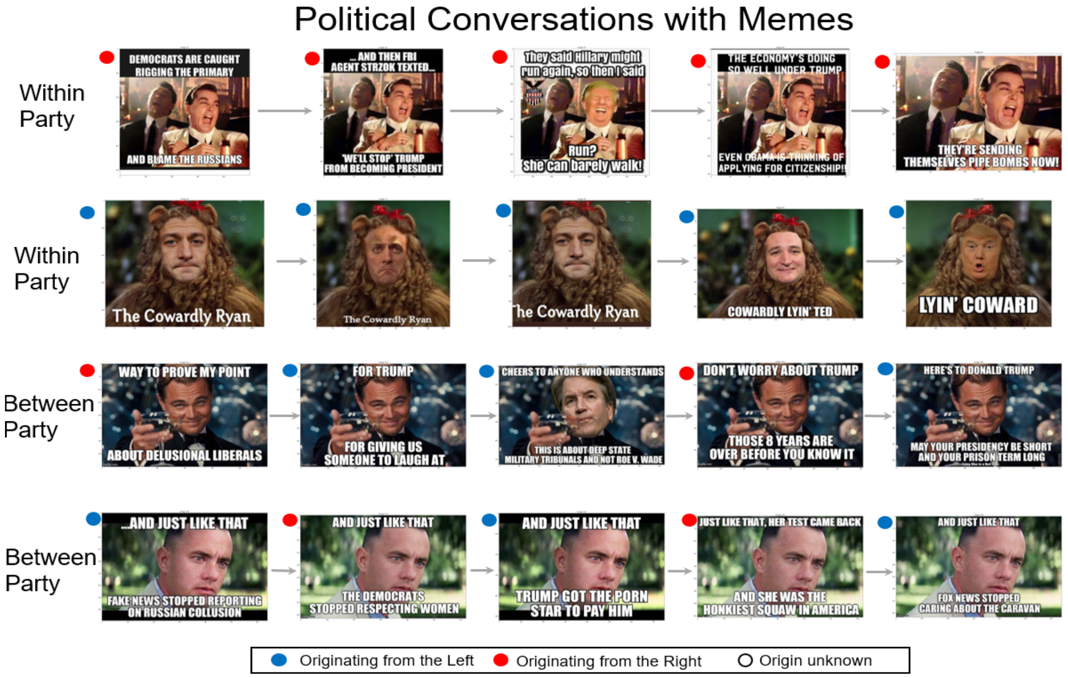

Political Conversations with Memes
Within Party
DEMOCRATS ARE CAUGHT RIGGING THE PRIMARY AND BLAME THE RUSSIANS
... AND THEN FBI AGENT STRZOK TEXTED... 'WE'LL STOP' TRUMP FROM BECOMING PRESIDENT
They said Hillary might run again, so then I said Run? She can barely walk!
THE ECONOMY'S DOING SO WELL UNDER TRUMP EVEN OBAMA IS THINKING OF APPLYING FOR CITIZENSHIP!!
THEY'RE SENDING THEMSELVES PIPE BOMBS NOW!
Within Party
The Cowardly Ryan
The Cowardly Ryan
The Cowardly Ryan
COWARDLY LYIN' TED
LYIN' COWARD
Between Party
WAY TO PROVE MY POINT ABOUT DELUSIONAL LIBERALS
FOR TRUMP FOR GIVING US SOMEONE TO LAUGH AT
CHEERS TO ANYONE WHO UNDERSTANDS THIS IS ABOUT DEEP STATE MILITARY TRIBUNALS AND NOT ROE V. WADE
DON'T WORRY ABOUT TRUMP THOSE 8 YEARS ARE OVER BEFORE YOU KNOW IT
HERE'S TO DONALD TRUMP MAY YOUR PRESIDENCY BE SHORT AND YOUR PRISON TERM LONG
Between Party
...AND JUST LIKE THAT FAKE NEWS STOPPED REPORTING ON RUSSIAN COLLUSION
AND JUST LIKE THAT THE DEMOCRATS STOPPED RESPECTING WOMEN
AND JUST LIKE THAT TRUMP GOT THE PORN STAR TO PAY HIM
JUST LIKE THAT, HER TEST CAME BACK AND SHE WAS THE HONKIEST SQUAW IN AMERICA
AND JUST LIKE THAT FOX NEWS STOPPED CARING ABOUT THE CARAVAN
● Originating from the Left ● Originating from the Right ○ Origin unknown

Figure 5.7: Political conversations within and between political left and political right.

memes as seen in Figure 5.7. In these images we see the cultural evolution that was originally envisioned by Richard Dawkins. We also see Linor Shifman's definition of memes play out as these meme images "lure extensive creative user engagement."

5.4.3 Results and Findings

Memes Usage in Election Events

Having identified memes thriving in the online conversation around these election events, we calculated descriptive statistics regarding memes and the accounts that share them. These descriptive statistics are provided in Table 5.3. In this table we make several observations that help us understand meme popularity and virality. First, we see that, although images are generally popular (high retweet/likes), memes are not. In both events, memes had fewer retweets and likes than other images, and in the US election memes had a shorter "life-span" on average. We hypothesize that the reason behind this is that attributed users do not want to associate their reputation with a controversial political meme and its message. For the same reasons that meme creators disassociate themselves from the memes they create, social media users, while influenced by memes, are hesitant to *like* or *retweet* them, especially polarizing political memes. If this is the case, then the virality of memes may not be due to normal social media activity (*like*, *share*, *retweet*), but rather occurs through the selection, retention, and mutation that Dawkins originally described. The memes mutate, carrying pieces of the original message, and are reintroduced in other corners of the Internet.

Algorithm 1 Memes Mutation Graph Algorithm

```
procedure GETMUTATIONGRAPH(Meme m)
    memes_graph ← new dictionary
    neighbours ← Get radius neighbours
    for $b_i$ in neighbours do
        if $b_i$ not in memes_graph then
            Add $b_i$ to memes_graph
    for $b_i$ in neighbours do
        if size(memes_graph) $\leq$ exit_condition then
            child_memes_graph ← $getMutationGraph$(b$_i$)
            Add child_memes_graph to memes_graph
    return memes_graph
```

We hypothesized that bots could be used to push memes on social media. Using the *bot-hunter* bot prediction tool [30] with a probability threshold of 0.6, we predicted the portion of accounts that have bot-like characteristics. In the Swedish data we found a slightly higher bot involvement with memes, but did not find this in the US election data. From this analysis we conclude that bot activity did not play an out-sized role in meme propagation for either of these events.

Additionally, we conducted face detection on the US election memes to find 18 prominent US politicians in the meme data. To do this we leveraged the open source face detection software created by Adam Geitgey and made available at [105], using a comparison threshold of 0.54. Using this face detection software, we found the distribution of memes by politician provided in Figure 5.8.

In Figure 5.9 we've plotted the posting or retweeting of meme images in the 2018 US Election by the political party of the candidate mentioned. Note that politicians and candidates are mentioned with both positive and negative memes. In this case, we see the highest volume of memes mentioning Democrats and Republicans associated with the time immediately after the Kavanaugh hearings.

Meme Propagation Across Platforms

Given the evolutionary and anonymous nature of memes, we hypothesized that memes propagate across the Internet differently than other viral content. Viral content is generally spread through the simple mechanisms of sharing, retweeting, liking, etc. Memes, as noted above, aren't liked or retweeted near as much as other media content. We believe that their propagation occurs more through their mutation and evolution, where one meme generates other creative works that emerge in other parts of the Internet. This would cause memes to 'hop' to more platforms and domains than normal images. While propagating to new corners of the Internet, however, the memes will undoubtedly morph, and this mutation is out of the hands and control of the original creators.

To assess this hypothesis, we sampled 5,000 meme images and 5,000 non-meme images from images associated with the 2018 US Mid-term elections. All images were political in semantic

Table 5.3: Descriptive Statistics about Internet Memes in Online Election Conversations.

	2018 Sweden Election			US Midterm Election		
Total Tweets	661K			62,034K		
Total Users	88K			2,695K		
Suspended/removed	1,616/2,302			41,901/47,349		
Total Images Shared	47K			4,446K		
Total Images Available	43K			4,037K		
	no image	meme	normal image	no image	meme	normal image
# Images Available		5K	38K		497K	3,539K
# of Unique Images		1.5K	10K		175K	951K
% of bot-like accounts	0.32	0.35	0.31	0.37	0.32	0.28
Life of tweet (hours)	0.51	0.60	0.59	21.80	16.02	22.87
Mean retweets	26	15	33	3,492	237	3,478
Mean Likes	0.84	1.50	2.03	15.96	24.42	65.48
User Median Followers	246	259	224	594	190	258
User Median Friends	348	401	340	857	375	407

and visual content. We then conducted a reverse image lookup or *web-detection* using the Google Vision API. This service provided us with links to matching and partially matching images on the Internet. The 5,000 meme images had 62,475 matching links associated with 9,536 unique domains. The 5,000 non-meme images had only 13,617 total links associated with only 4,731 unique domain names. The memes therefore were connected to roughly 4 times the number of links and twice the number of domains when compared to non-meme images, supporting the hypothesis that memes propagate to more corners of the Internet than other types of media.

5.4.4 Comparison to Past Methods

In our section looking at *related works*, we noted several research efforts that leverage meme templates. These efforts include multi-model efforts by Dudley et al. [85] and meme evolution effort by Zannettou et al. [264]. While Dubey uses this technique for virality prediction and clustering, we primarily want to compare their approach to meme hunter for the task of image retrieval (i.e. extracting all meme images in a given social media stream). The primary limitation to their work is that it is constrained to identify memes found on sites like Memegenerator or Quickmeme. As we illustrate below, this approach, while generating high precision, finds very few of the total memes in election-related social media streams (low recall). The Meme-Hunter approach that we propose, while limited to only two types of memes, typically finds at least 8 times more memes in election-related social media streams as approaches constrained by meme templates.

To evaluate both methods, we randomly sampled 1,050 images from both the Swedish elec-

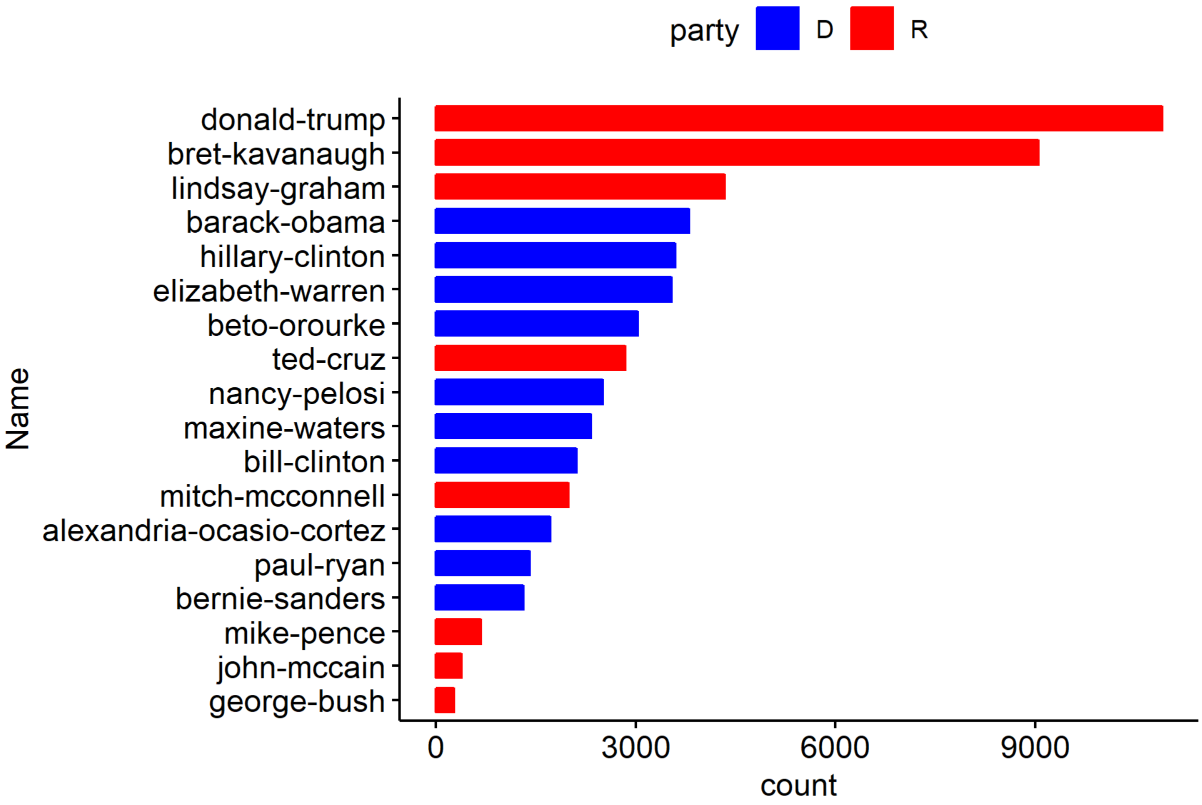


Figure 5.8: Memes by Politician (identified by Face Detection).

tion event and 1,050 images from the 2018 Midterm election stream. We then manually labeled any image that could be construed as an Internet meme as defined by Dawkins and Schifman. We then ran our Meme-Hunter approach and compared this to a template-based approach.

To replicate a template-based approach, we collected 39,112 meme templates from the Meme Generator web application found at `https://imgflip.com/memegenerator`. This included most of the popular and even less popular meme templates used, to include meme templates associated with politicians and world leaders. We then used perceptual hashing (phash) to identify any image in the test image set that used one of the meme templates. Positive matches were determined by those hashes that required less than 10 substitutions in a Hamming distance comparison. Positive matches were then considered memes.

Meme hunter was applied with unimodal machine vision models as well as multi-modal models as indicated in Table 5.4. In this comparison we see that, while template-based approaches offer high accuracy and precision, the recall in both election-based streams is only approximately 5%. In these very dynamic political dialogues, many images that are construed as memes are not yet in the template databases. This means that using template-based methods will only find 5% of the memes in these streams. The Meme Hunter approach, while offering slightly lower accuracy and precision, is able to find 8× more memes, with the InceptionV3 unimodal model and all multi-modal models providing the highest performance across all metrics. We see that, in regard to the accuracy metric, multi-modal consistently outperforms unimodal models. The top models

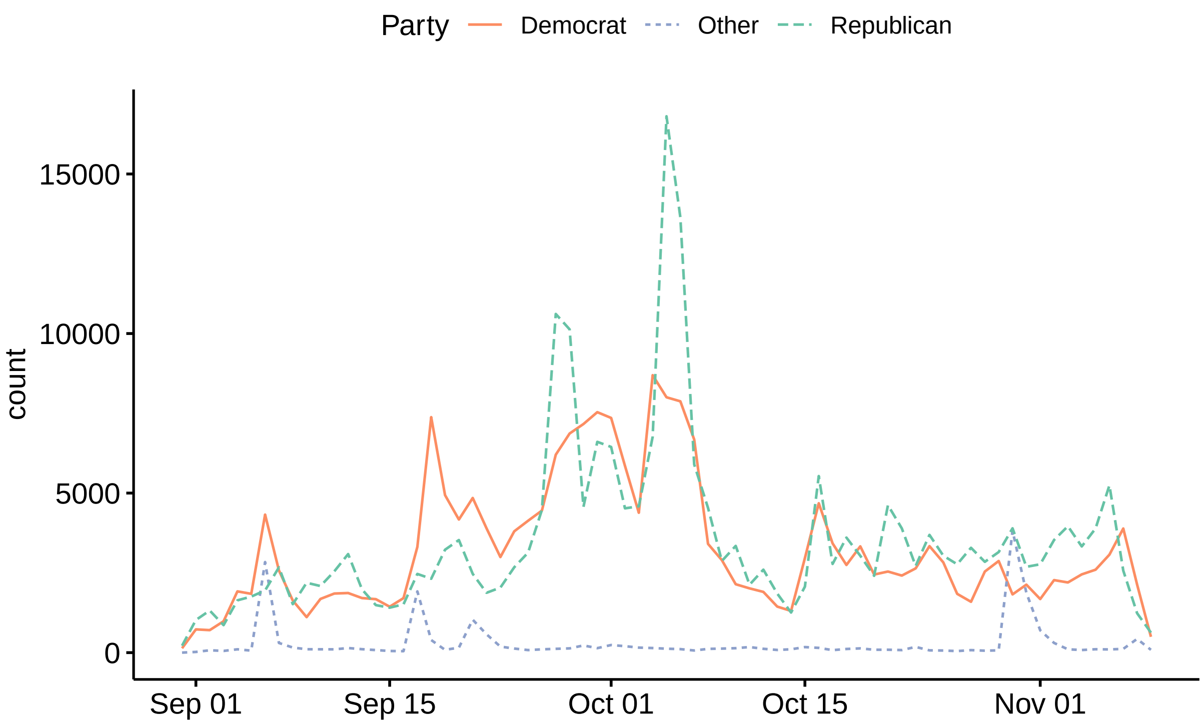


Figure 5.9: Memes (both positive and negative) by Political Party of Candidate mentioned.

using the Meme-Hunter DNN approach find approximately 50% of the images in both streams.

In this comparison we also want to comment on the lower performance of Meme-Hunter in the US Midterm stream compared to the Swedish election stream. This is the result of more sophisticated memes being used in the US election stream, some of which are elaborate photo editing work flows and contain no text. Others contain vertical text or specially placed text. Meme-Hunter will struggle to positively identify these more sophisticated memes.

5.5 Conclusion

In this chapter we present a method for using deep learning to classify memes and graph learning to cluster them into their evolutionary “families”. Additionally, these models were used to analyze meme usage inside two large democratic election events. We found that Meme-Hunter provided at least 8 times higher recall than template-based methods and that graph learning can capture the overall structure of the evolutionary tree. Having identified memes in large election events, we found evidence that memes are liked and retweeted less, but families of memes ‘hop’ platforms and travel to more locations of the Internet than regular images. This indicates that memes do not propagate across social media and the Internet in the same way as other viral content. While our research primarily analyzed memes extracted from Twitter, the Meme-Hunter model can be used on images extracted from any source or platform.

Table 5.4: Comparing Meme-Hunter to meme template-based approaches to find memes in social media streams.

	Sweden				US Midterms			
Model	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
template-based	**0.872**	0.107	**0.727**	0.058	0.795	0.100	**0.667**	0.054
VGG18	0.809	0.437	0.358	0.561	0.771	0.348	0.435	0.290
ResNet18	0.846	0.464	0.429	0.504	0.806	0.430	0.562	0.348
Inception V3	0.820	0.488	0.391	**0.647**	0.807	**0.494**	0.550	**0.448**
Vision + Text	0.865	0.510	0.490	0.532	**0.815**	0.455	0.600	0.367
Vision + Text + Face	0.858	**0.511**	0.470	0.561	0.812	0.439	0.592	0.348

The organic and evolutionary nature of memes has caused some nation states to ban them [168], while encouraging other nations to leverage them as part of elaborate propaganda operations [115]. The countries that ban them do so largely because memes evolve outside of the control of the state and because image memes can be difficult to trace [1]. Those countries that leverage them for information warfare do so for the exact same reasons. We hope that our proposed methods to study memes would provide more possibilities to trace memes for good causes.

Chapter 6

Developing Bot Labels

While substantial research has focused on social bot classification, less effort has focused on repeatable computational bot characterization. Bot characterization provides the account forensics necessary to confirm non-human behavior, fingerprint and identify the state or non-state actor behind the operations, and possibly find evidence of intended target and intended effect. Together these provide an important step in the social cybersecurity workflow. The focus of bot-labels is to develop a computational pipeline for applying non-exclusive labels to suspicious accounts to aid researchers and analysts in characterizing these accounts and the operations they support and enable.

6.1 Introduction

Significant research over the last decade has developed machine learning and computational methods to classify and otherwise detect bots in social media. These efforts enable the first and very important step in social cyber security, namely identifying the threat. Far less research, however, has focused on repeatable computational methods to characterize this threat. The goal of this paper is to provide a repeatable and general purpose computational tool to characterize social cyber security threats by applying badges or labels in ways that could be leveraged by data scientists conducting deep dives in Python code or on the back end of a social cyber security tool used by analysts, journalists, and industry. These labels and their associated methods can be used together in a concerted pipeline or independently as stand-alone analysis.

The majority of social media bot classification methods render a binary classification: *bot* or *not bot*. Within social cyber security workflows, this is the necessary first step in threat 'triage'. In most streams and conversations, however, this step returns a subset of data that is still too large for a deep dive. A few research efforts have attempted to solve this problem by leveraging a multi-class machine learning approach, placing accounts into multiple classification 'buckets' instead of just two [65, 134]. These buckets, however, are mutually exclusive...an account can only reside in a single classification bucket. Additionally, the authors generally don't have more than three or four buckets, meaning that these are still very broad and general categorization.

The overarching goal in bot-labeler is to apply multiple non-exclusive labels to social cyber security threats. For example, an account can be labeled as an *amplification* account with *honey pot* profile or as a *spam* account with *high volume*. In applying these labels, analysts and researchers gain insight into the likely purpose of the account. These labels can then be aggregated in order to characterize the larger (dis)information operation. Finally, these labels provide tokens that enable analysts to query and subset large streams ("show me all *amplification accounts* that have *high volume*").

In choosing labels, we attempt to automate the mental heuristics that expert social cyber security analysts use when they analyze an account. These mental heuristics have been shared in multiple handbooks, articles, and reports [133, 182]. Automating some labels requires simple heuristic algorithms, while other labels rely on more complicated deep learning algorithms. By packaging these algorithms in a single Python package, we provide a toolbox of individual analysis tools as well as a pipeline to run all the analysis on accounts of interest.

At a higher level the labels are aggregated into *profile* labels, *content* labels, and *network* labels. The *profile* labels help characterize the profile of the accounts, analyzing the name, description, profile image, etc. *Content* labels assist in characterizing the timing and nature of the content that the account is posting. Finally, *network* labels are applied to the surrounding network (friends, followers, mentions).

Having developed the computational methods to apply overlapping labels to social cyber security threats, we then evaluate these labels by running them on two relevant data sets. The first is likely bot accounts that have been actively participating in the Twitter discussion around the Canada 2019 National Elections. The second data set is a scam related social bot-net often called the "McAfee" bots.

By developing these overlapping labels and computational methods to apply them, this effort contributes an essential piece to the social cyber security workflow while enhancing our understanding of the social cyber security threat and how it interacts with and attempts to manipulate the core beliefs and values held by a society.

6.2 Literature Review

Multiple studies over the last decade have conducted the needed first step of bot classification. This step triages the social media stream and identifies accounts that are likely nuisance or malicious. Several models are commonly used today, including supervised machine learning models Botometer (previously BotOrNot) [74] and Bot-Hunter [30], unsupervised machine learning models like the Debot algorithm [60], graph based models like Sybil-Guard [262], and anomaly detection models like Bot-Walk [173]. Some models attempt to generalize (Botometer and Bot-Hunter), others are focused on a specific type of bot, like Debot's focus on correlated accounts. All of the models help identify part of the threat.

Many studies have characterized bot and malicious accounts in specific events. Researchers have looked at disinformation in the 2012 Korean Presidential Elections [137], 2016 US Elections [39, 193, 265], the French National Elections [92], the Brexit Vote [130] as well as other events. In most cases these provide exploratory data analysis through summary statistics and descriptive visualization, at times preceded by bot detection/classification. In almost all cases this analysis is tailored to the event in question and does not provide general and repeatable characterization.

Recently investigative journalism has contributed several salient articles to help characterize bots. The most prominent of these was the New York Times Article “Follower Factory” which opened many people’s eyes to the bot problem and provided novel analytic and visual techniques to help characterize bot behavior [68]. Other prominent works were produced by the Pew Research Center [254], the RAND Corporation [124] and the Digital Forensic Labs [182].

Several other works are closely related to bot-labeler. These include ‘BotCamp’ which clusters bots into a given political campaign and then categorizes their interaction [4]. Additionally, VASSL uses some account characteristics to support a visual tool that enables manual bot labeling tasks [138]. Finally, the ‘TruthNest’ project was a European Union funded REVEAL project designed to provide people in the European Union with a tool to analyze Twitter accounts. It provides the most robust account characterization pipeline and interface that we’ve seen yet, but access is severely limited by a pay-wall.

Today most social cybersecurity research focused on characterizing social media bots will focus on a niche research effort on a single known campaign plan, such as the 2016 US Elections [4], the IRA meddling of US elections [20, 83], or the Chinese information campaign focused on Hong Kong [238]. This is often conducted by an experienced social cybersecurity expert with a tailored workflow. While providing many insights for the given event, the approach often does not provide a repeatable, generalizable workflow that is accessible to the non-expert. Our goal is to provide this generalizable and repeatable workflow for bot characterization using non-exclusive bot labels.

6.3 Characterization in Social Cyber Security Workflow

In Figure 6.1 we review a typical social cybersecurity workflow. This workflow begins with filtering social media streams using content, network, or geographic filters. Once a curated stream is built, initial supervised machine learning algorithms are the first “pass” over the data and help to quickly identify the majority of the malicious actors. Classification, however, often only renders a binary result (bot/not bot, troll/not troll). The next step is to characterize and better understand these actors. This is where *Bot-Labeler* comes in.

Note that the results of the characterization are an important input for the rest of the social cybersecurity workflow. The results can be aggregated to understand the campaign and campaign desired effects. The results help characterize the message. Importantly, they also provide forensic evidence that can help with identifying the perpetrator (i.e. attribution). Finally, certain bot characteristics help identify the target audience. All of this provides intermediate data that can be used to measure the impact of the campaign.

6.4 Data

We will apply and analyze labels on two types of bot data. The first data set is bots involved in an election event, the second is bots involved in a deliberate online scam. The nature of the accounts

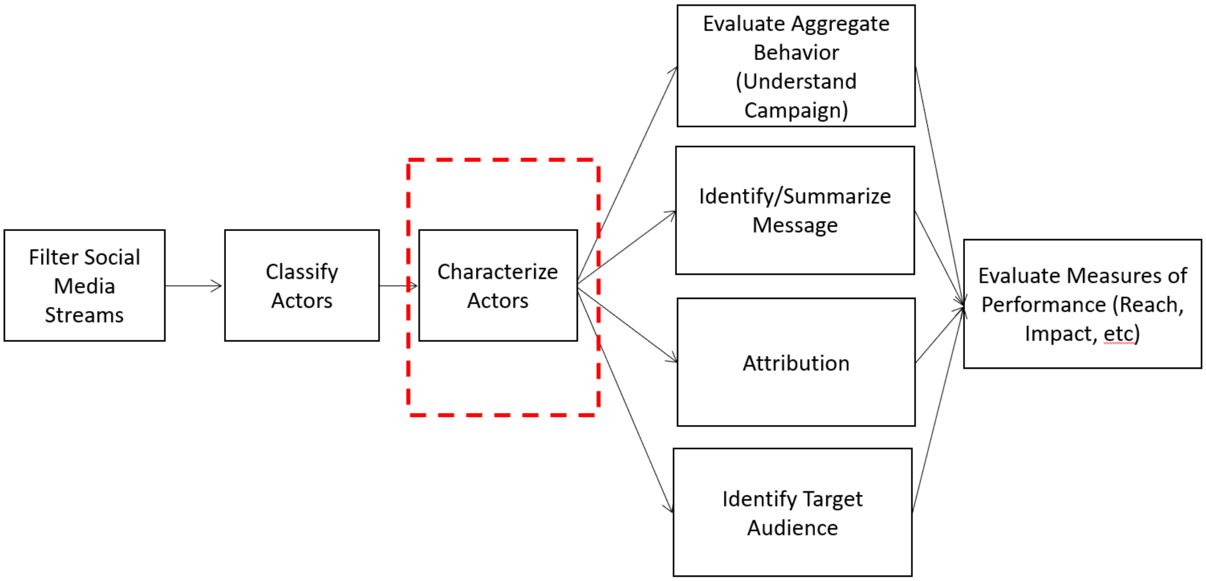

Filter Social Media Streams
Classify Actors
Characterize Actors
Evaluate Aggregate Behavior (Understand Campaign)
Identify/Summarize Message
Attribution
Identify Target Audience
Evaluate Measures of Performance (Reach, Impact, etc)

Figure 6.1: Social Cyber Security Workflow

and distribution of labels for these will demonstrate the differences in bot characteristics across bot genres.

6.4.1 Canadian 2019 National Elections

The *canada* data set includes bots that were detected in the online Twitter discussion surrounding the Canadian 2019 National Elections. The initial data was collected by filtering key Canadian election hashtags from the Twitter Streaming API. After collecting the Canadian election stream, we applied a supervised bot detection algorithm to identify likely bot accounts. This supervised model serves as an initial triage of the data. The resulting subset of tweets and accounts was still too large to conduct a deep dive. This is where *Bot-Labeler* would come in to create mutually overlapping labels for these accounts allowing an analyst or investigative journalist to query labels of interest.

The following hashtags were used to filter the stream: #TrudeauMustGo, #TeamTrudeau, trudeau, #Election2019, #ItsOurVote, #elxn43, #chooseforward, #onpoli, #ItsOurVote, #lpc, #ndp, #cpc, #gpc, #NotAbot, #cdnpoli, #ButtsMustGo, #LavScam, #LiberalsMustGo, BlocQuebecois, #blocqc, #cccr2019, #NoTMX, #TMX, #TransMountain, #scheer, #dougford, #fordcutshurt, #fordisfailing. The hashtags were identified by starting with a few seed hashtags, and then every few weeks checking the stream for new emerging hashtags and adding these to the filter. While not covering the full spectrum of politics in Canada, we found that this collection of terms seemed to give a good overall representation of the Canadian political conversation. It should be noted that throughout this collection we found that the Canadian political conversation and the US political conversation were closely intermingled. This collection occurred between 20 July and 6 November 2019, and produced 16,784,400 tweets from 1,303,761 unique users as shown in Figure 6.2.

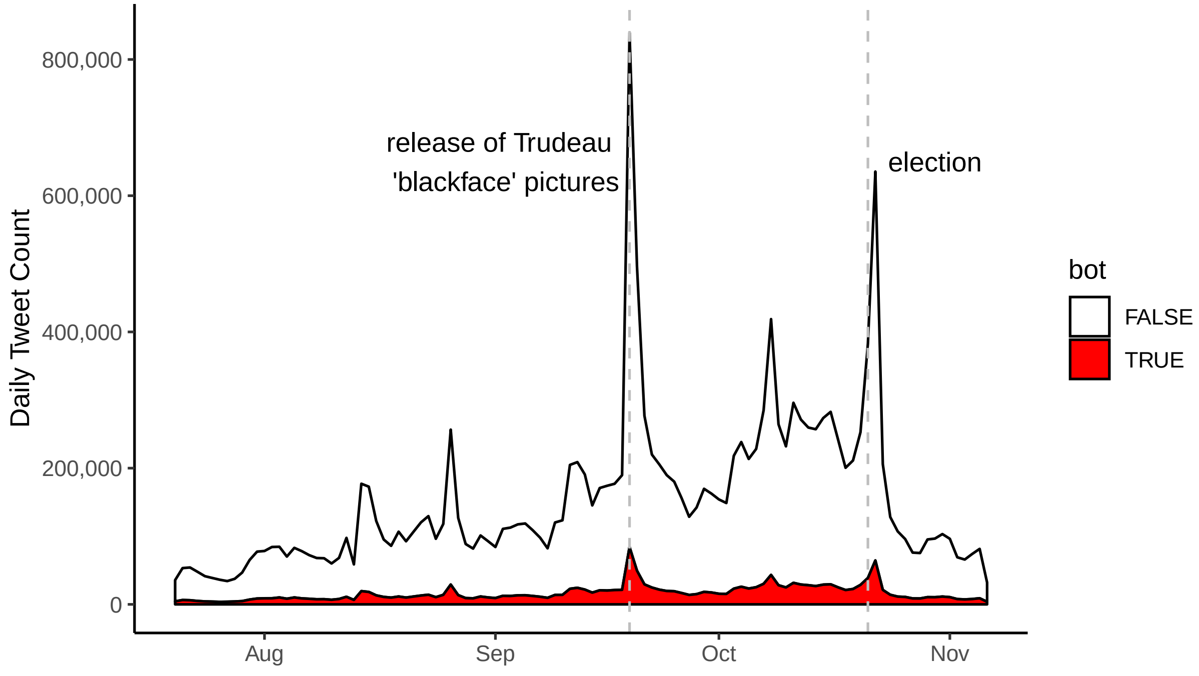


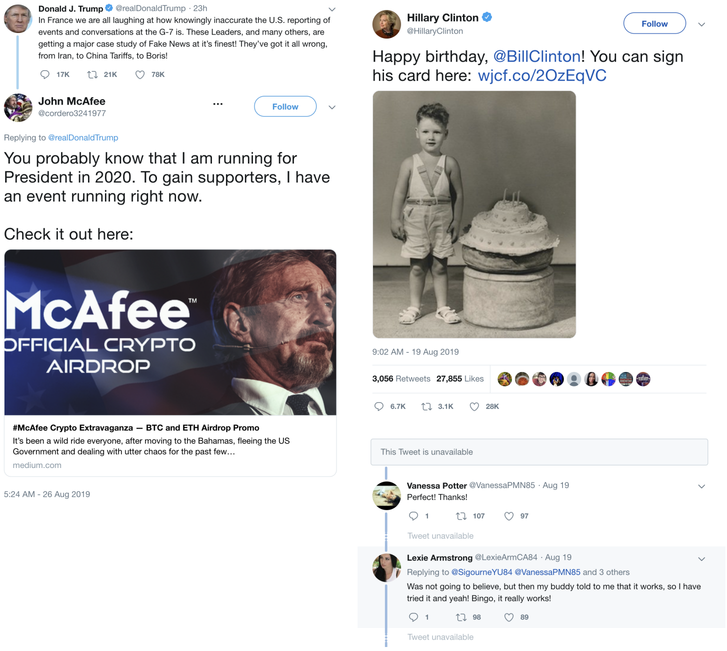
Figure 6.2: Daily Tweets related to the Canadian 2019 election

Having collected the Canadian election stream, we next classified the accounts in the stream using the bot-hunter Tier 1 algorithm with the threshold set at 0.65. The Bot-Hunter tool [30] is a supervised machine learning algorithm trained on features extracted from both the *user* object and *tweet* object returned by the Twitter Application Programming Interface (API). The Bot-Hunter Tier 1 model was trained on sophisticated bots that harassed NATO and the Digital Forensic Labs in 2017. The Bot-Hunter Tier 1 model was designed to conduct classification at scale on existing Twitter JSON, which made sense for our problem set.

The resulting data set contained 1,815,027 tweets produced by 123,042 unique accounts that were labeled as a bot with a probability threshold of 0.65.

6.4.2 McAfee Bot Scam

The second data set consists of a sophisticated set of coordinated bots that were documented in a Bitcoin scam [2]. The bots were programmed to automatically reply to prominent US politicians, including Donald Trump and Hillary Clinton, and were routinely the first reply to a Tweet produced by either of these two prominent US Politicians. The bots advertised a Bitcoin scam, and then had a consortium of other bots in their network that would reply claiming to have successfully participated in the advertised scam. The tweets advertised the scam by sending the intended victim to a page on Medium that detailed the scam. Even as Twitter and Medium both targeted this network, it seemed resilient, creating many more accounts and multiple Medium pages with the same content.



Donald J. Trump @realDonaldTrump · 23h
In France we are all laughing at how knowingly inaccurate the U.S. reporting of events and conversations at the G-7 is. These Leaders, and many others, are getting a major case study of Fake News at it's finest! They've got it all wrong, from Iran, to China Tariffs, to Boris!
17K 21K 78K

John McAfee
@cordero3241977
Replying to @realDonaldTrump
You probably know that I am running for President in 2020. To gain supporters, I have an event running right now.

Check it out here:

#McAfee Crypto Extravaganza — BTC and ETH Airdrop Promo
It's been a wild ride everyone, after moving to the Bahamas, fleeing the US Government and dealing with utter chaos for the past few…
medium.com

5:24 AM - 26 Aug 2019

Hillary Clinton
@HillaryClinton
Happy birthday, @BillClinton! You can sign his card here: wjcf.co/2OzEqVC

9:02 AM - 19 Aug 2019
3,056 Retweets 27,855 Likes
6.7K 3.1K 28K

This Tweet is unavailable

Vanessa Potter @VanessaPMN85 · Aug 19
Perfect! Thanks!
1 107 97

Tweet unavailable

Lexie Armstrong @LexieArmCA84 · Aug 19
Replying to @SigourneYU84 @VanessaPMN85 and 3 others
Was not going to believe, but then my buddy told to me that it works, so I have tried it and yeah! Bingo, it really works!
1 98 89

Tweet unavailable

Figure 6.3: The bots would immediately reply to prominent tweets from Donald Trump, Hilary Clinton, Elon Musk, and other prominent voices. The tweets would appear to be from John McAfee, and would direct the victims to multiple Medium posts that would attempt to scam the victim. An important part of the scam is social validation though multiple users from the bot net replying and indicating that they had been able to receive the Bitcoin. Note that in the Hilary Clinton thread above, the original scam post has been deleted, but the replies are still present.

The accounts in this network appeared to have some distinguishing characteristics. All account screen names appear randomly generated by concatenating an English male or female name with two uppercase letters and two numeric digits. An example is seen in Figure 6.3 is **LexieArmCA84**. Many accounts also use risqué pictures in their profile image and posted images to increase followers from a certain demographic. This characteristic is explored more later.

Unlike the *canada election* data, the *McAffee bot* data was collected using snowball sampling on Twitter. We began identifying the McAffee scam tweets, and recording the sender as well as the 4-8 accounts that automatically replied claiming to validate the advertised offer. We then recursively collected followers of these seed accounts, filtered by distinguishing characteristics (similar screen name template tweeting about Bitcoin), which we found were likely part of this bot scam network. Given this data collection method, we identified 1,006 users as part of the McAffee bot scam. We used this as our second data set to characterize with labels.

The Mcafee and Canada data provide us with two malicious data sets with diverging motivations and intents. We will demonstrate the bot-label methodology in profiling these two datasets.

Some basic statistics for both data sets is provided in Table 6.1.

Table 6.1: Summary Data

Metric	Canadian Elections Data	McAfee Data
Tweets	1,815,027	103,954
Users	123,042	1,006
% of replies	7.5%	4.9%
% of retweets	60.6%	32.1%
% of quotes	28.0%	2.5%
% of original	3.9%	60.5%
mentions % $(0 \vert 1 \vert > 1)$	8% | 66% | 26%	62% | 33% | 5%
hashtags % $(0 \vert 1 \vert > 1)$	75% | 13% | 12%	75% | 8% | 17%
median/mean friends	868 | 1360	443 | 488
median/mean followers	197 | 2121	58 | 62
prominent hashtags	cdnpoli, elxn43, TrudeauMustGo, Trudeau, CPC	gameinsight, androidgames, android, investing, Topix

6.5 bot_labeler

To accomplish the task of characterizing accounts, we chose to apply non-exclusive labels to an account. This deviates slightly from approaches that attempt to classify accounts into multiple exclusive categories. Having studied bot genres in multiple settings and events, our team has found that the categories overlap, and exclusive categories would be both constraining and likely too specific. The labels that we apply are determined by various algorithms that we've developed or in some cases where we've leveraged available open source algorithms.

The labels fall into three overarching categories: 1) profile labels, 2) content labels, and 3) network labels. Profile labels help us understand account features and metadata, content labels help us understand the narrative that is generated with tweets and replies or propagated with retweets, and network labels help us understand the type of accounts that link to the account in question.

6.5.1 Profile Labels

The labels below address characteristics of the account profile such as the description, image, name, and other characteristics. These help to identify if certain anomalies are occurring with the account. Below we will walk through each of the labels, indicate when the label is returned true, and describe any associated metadata that bot_labeler returns with the respective label.

Default Profile

Some bots accounts don't change their default image in an effort to create anonymity. This is often used for intimidation bots and other bots that don't need to look like a real person. This label is positive if the account in question has not changed their default profile image. This label does not contain any additional metadata.

Gender Mismatch

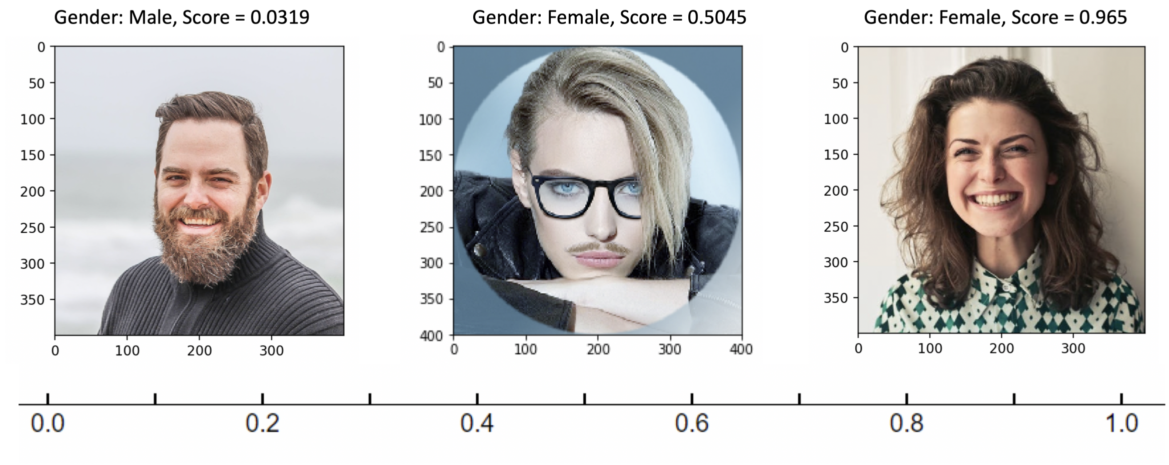

Gender: Male, Score = 0.0319
Gender: Female, Score = 0.5045
Gender: Female, Score = 0.965
0
50
100
150
200
250
300
350
400
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0

Figure 6.4: Demonstrating Gender Prediction Performance

Some bots found by our team as well as other researchers [182] have a likely gender mismatch between the profile name and the profile image. For example the profile image shows a female while the name is male, or vice versa. As bot armies are at times automatically populated, care isn't shown to ensure gender matching. To check for gender match, we employed two available open source tools for predicting gender based on image and name, respectively. Our model uses the Pyagender open source package [1] to estimate the age and gender of each face in the profile image, and then uses the gender_guesser python package [2] to estimate the gender of the first tokenized word in the user name string. Pyagender performance is provided in Figure 6.4. For each face in the image, Pyagender renders gender ('male', 'female') and age, whereas gender-guesser renders the following possible labels: *unknown*, *androgynous*, *male*, *female*, *mostly male*, *mostly female*. Bot-labeler only renders gender mismatch as true if the name is clearly labeled male or female and that respective gender is not found in any marked faces of the profile image. The *gender mismatch* label also provides the prediction and score for the image gender prediction and the prediction for the name gender. Examples of gender mismatch are provided in Figure 6.5.

[1] https://pypi.org/project/py-agender/
[2] https://pypi.org/project/gender-guesser/

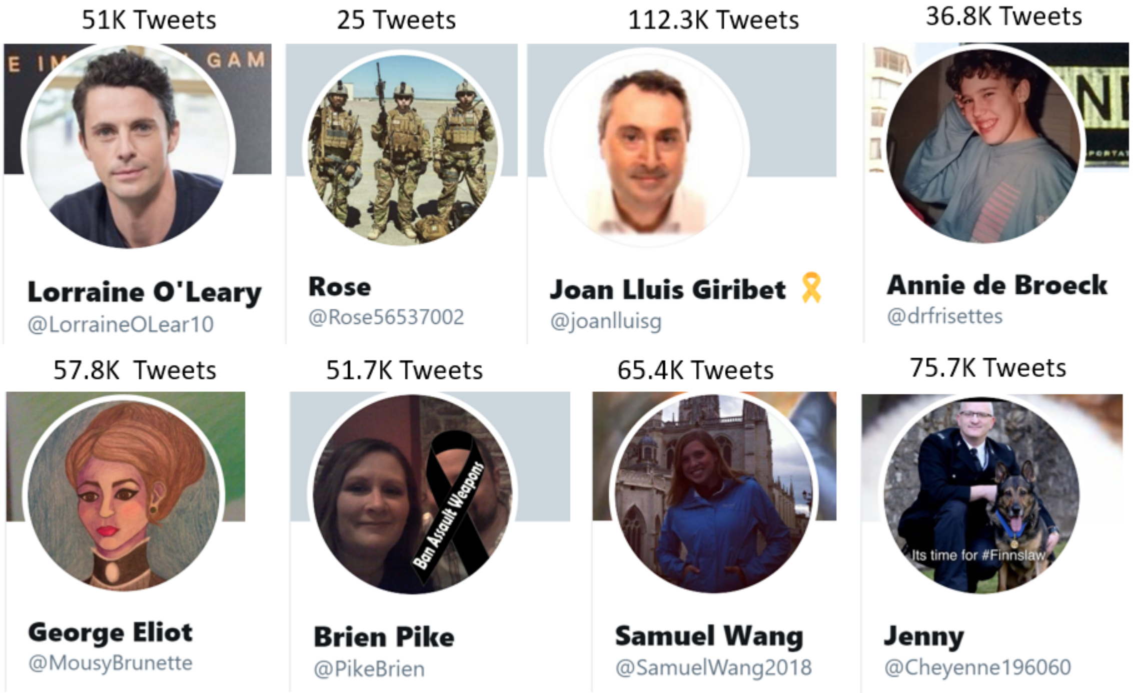


Figure 6.5: Bots from both the left and right of the Canadian Elections using gender mismatched accounts

Honey Trap Account

Since antiquity spies and other shadow organizations have used sex and love to obtain information or action from their enemies. In World War 1 a Dutch exotic dancer named Mata Hari spied for the Germans, and in the cold war male East German "Romeo" spies used "love" and "lies" to obtain intelligence for East Germany and the Soviet Union [7]. In the world of espionage, these agents are known as "honey traps".

On social media accounts with risqué images and content are used alongside information operations to attract certain audiences [27] and deploy cyber hacks [250]. Several state owned propaganda sites, particularly Sputnik International, regularly interlace risqué images and content in order to attract viewership from specific audiences. Note that adult content accounts focus almost exclusively on adult content, while "honey trap" accounts intermix risqué content with their target message or operation.

In `bot_labeler` we use Yahoo's open source 'Not Safe for Work' deep learning model [163] to predict whether a profile image is beyond the level of risqué of a normal Twitter user, and is therefore either an overt adult content page or a 'honey trap'. Yahoo's NSFW model renders a probability that ranges from $0-1$ (this value is provided in the label meta-data). While definitions of 'risqué' and 'NSFW' are subjective, we found that the Yahoo model single metric proved very effective in our task. Using Precision-Recall curves we determined that if honey-pot accounts are expected (as in the McAfee data), the threshold should be between 0.05 and 0.15. If honey pot accounts are expected to be a rare event (as in the Canadian election and most election events), then the threshold should be 0.3 in order to minimize False Positives. This algorithm is demonstrated in Figure 6.6.

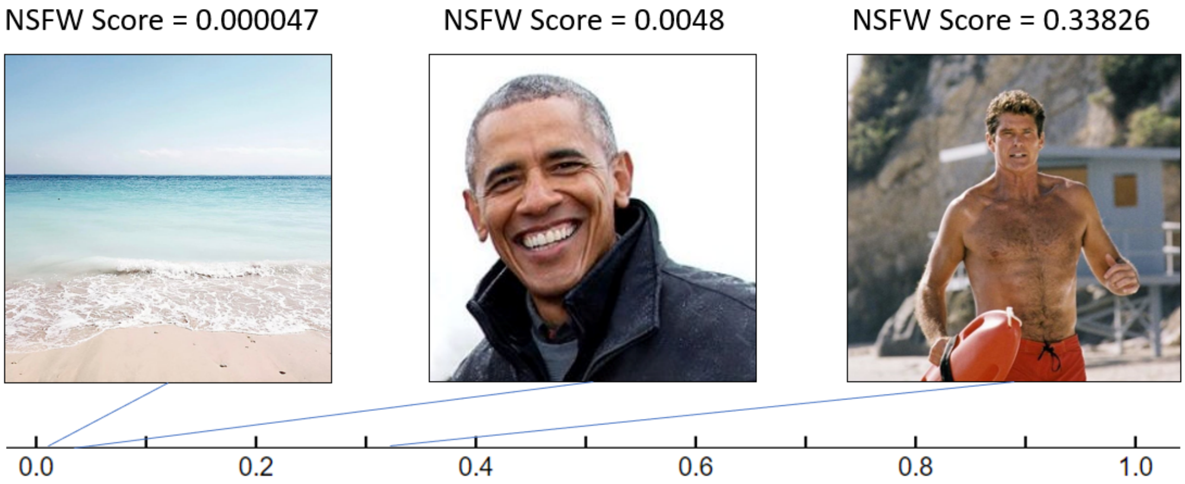


Figure 6.6: Example NSFW Scores from Yahoo NSFW model

6.5.2 Content Labels

These labels identify characteristics of the content and the rate at which the content is produced. In doing so, they help ascertain the narrative as well as certain types of information forms of maneuver.

Amplification

Many Twitter bots are used to amplify specific voices online. This is easy to automate with retweets. A given bot or bot army will be provided with a list of accounts or topics to amplify, and then simply monitor and retweet the designated accounts or topics. A common method that social cybersecurity practitioners use to find *amplification* accounts is to find accounts that have a high proportion of retweets. In `bot_labeler`, we automate this by collecting the last 400 tweets produced by an account, and measuring the proportion of retweets. The amplification label is returned True if the proportion is greater than 70%. The raw proportion is returned in the metadata.

High Volume

Bots by their very nature are able to scale better than humans, conducting social media transactions at the speed of algorithms. High volume therefore provides a strong indicator of bot activity. As this volume increases, at a certain point we begin to wonder if this is truly a human. Not many people have the ability to remain active on social media as focused and steadfast as these high volume bots demonstrate. For the high volume computation, we use the last 400 tweets to find an average number of tweets per day. If this number is more than 50, we flag the account for high volume and provide this value in the metadata. The maximum mean daily tweets found in the McAfee data is 147 tweets per day. The maximum mean daily tweets found in the Canada data set was 200 tweets per day.

Spam

There are lots of definitions and types of spam. In our case we apply this label to a narrow subset of bots that tend to add numerous mentions to their tweets in order to gain the attention of the mentioned accounts, in order to amplify certain accounts, and in order to manipulate how Twitter prioritizes content and accounts in the Twitter feed. `Bot_labeler` applies the *spam* label if the median number of mentions in the last 400 tweets is 3 or more.

Random Language

Some bots retweet random content. This is sometimes done by hobby bots, but is also a tactic used by more malicious bot architects. If a bot retweets so much random content, it will be hard for anyone looking at the account to determine a more sinister purpose for the account. Buried in the random content is a signal that is trying to amplify specific content and actors. The random content is therefore used to mask this signal. The random language label is applied to any account that regularly uses more than 5 languages (5 languages have more than 1 tweet/retweet). For example, one account in the Canadian Election Data used 23 languages in the last 400 tweets.

High Memes

The internet meme has emerged as a powerful tool in information operations. The meme often combines a humorous image with culturally aware text to help a message resonate with a target audience. Often these memes are created to appeal to existing biases. The nature of meme evolution also provides anonymity and a natural organic way for quality memes to propagate across the internet. The authors have previously developed `meme-hunter`, a multi-modal deep learning solution (with meme specific Optical Character Recognition) that can classify images as a meme [38]. In `bot_labeler` we first collect the last 400 tweets by the evaluated accounts, then scrape all images shared in these tweets, and finally run `meme-hunter` in order to determine how many of these images are likely memes. If the account has more than 20% of the images classified as memes, then we apply the *high meme* label. Examples of memes in the Canadian election data is seen in Figure 6.7.

Duplicate Pictures

Bots are often programmed to scrape from a given source (internet site, social media account, etc) and post Tweets, which can result in a bot tweeting the same text and multimedia multiple times if the source hasn't changed. One way to identify bot activity is to scan an account's timeline to identify duplicated content, which is easiest to identify with duplicate pictures. In `bot_labeler`, positive labels for duplicate images is indicated if any image within the last 400

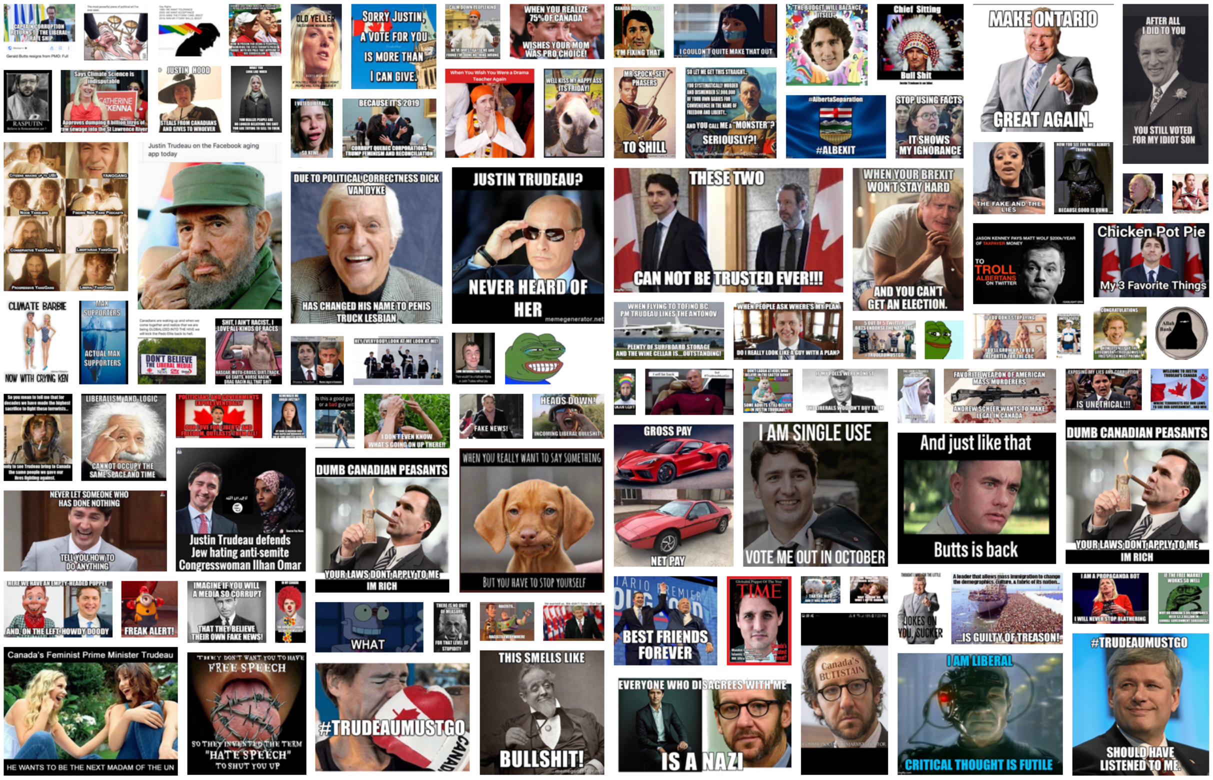

RASPUTIN
JUSTIN HOOD
CATHERINE MCKENNA
OLD YELLER
SORRY JUSTIN, A VOTE FOR YOU IS MORE THAN I CAN GIVE.
BECAUSE IT'S 2019
WHEN YOU REALIZE 75% OF CANADA WISHES YOUR MOM WAS PRO CHOICE!
I'M FIXING THAT
I COULDN'T QUITE MAKE THAT OUT
TO SHILL
#AlbertaSeparation
#ALBEXIT
Chief Sitting Bull Shit
STOP USING FACTS IT SHOWS MY IGNORANCE
MAKE ONTARIO GREAT AGAIN.
AFTER ALL I DID TO YOU YOU STILL VOTED FOR MY IDIOT SON
THE FAKE AND THE LIES
Justin Trudeau on the Facebook aging app today
DUE TO POLITICAL CORRECTNESS DICK VAN DYKE HAS CHANGED HIS NAME TO PENIS TRUCK LESBIAN
JUSTIN TRUDEAU? NEVER HEARD OF HER
THESE TWO CAN NOT BE TRUSTED EVER!!!
WHEN YOUR BREXIT WON'T STAY HARD AND YOU CAN'T GET AN ELECTION.
TROLL ALBERTANS ON TWITTER
Chicken Pot Pie My 3 Favorite Things
CLIMATE BARBIE
NOW WITH CRYING KEN
ACTUAL MAX SUPPORTERS
DON'T BELIEVE THE LIBERAL MEDIA
LIBERALISM AND LOGIC CANNOT OCCUPY THE SAME SPACE AND TIME
I DON'T EVEN KNOW WHAT'S GOING ON UP THERE!!
FAKE NEWS!
HEADS DOWN!
IS UNETHICAL!!!
GROSS PAY
NET PAY
I AM SINGLE USE VOTE ME OUT IN OCTOBER
And just like that Butts is back
DUMB CANADIAN PEASANTS YOUR LAWS DONT APPLY TO ME IM RICH
NEVER LET SOMEONE WHO HAS DONE NOTHING TELL YOU HOW TO DO ANYTHING
Justin Trudeau defends Jew hating anti-semite Congresswoman Ilhan Omar
DUMB CANADIAN PEASANTS YOUR LAWS DONT APPLY TO ME IM RICH
WHEN YOU REALLY WANT TO SAY SOMETHING BUT YOU HAVE TO STOP YOURSELF
AND, ON THE LEFT HOWDY DOODY
FREAK ALERT!
IMAGINE IF YOU WILL A MEDIA SO CORRUPT THAT THEY BELIEVE THEIR OWN FAKE NEWS!
WHAT
BEST FRIENDS FOREVER
TIME
JOKES ON YOU, SUCKER
IS GUILTY OF TREASON!
#TRUDEAUMUSTGO
Canada's Feminist Prime Minister Trudeau
HE WANTS TO BE THE NEXT MADAM OF THE UN
THEY DON'T WANT YOU TO HAVE FREE SPEECH SO THEY INVENTED THE TERM "HATE SPEECH" TO SHUT YOU UP
#TRUDEAUMUSTGO
THIS SMELLS LIKE BULLSHIT!
EVERYONE WHO DISAGREES WITH ME IS A NAZI
Canada's BUTTS
I AM LIBERAL CRITICAL THOUGHT IS FUTILE
SHOULD HAVE LISTENED TO ME

Figure 6.7: Examples of memes found in the 2019 Canadian election stream

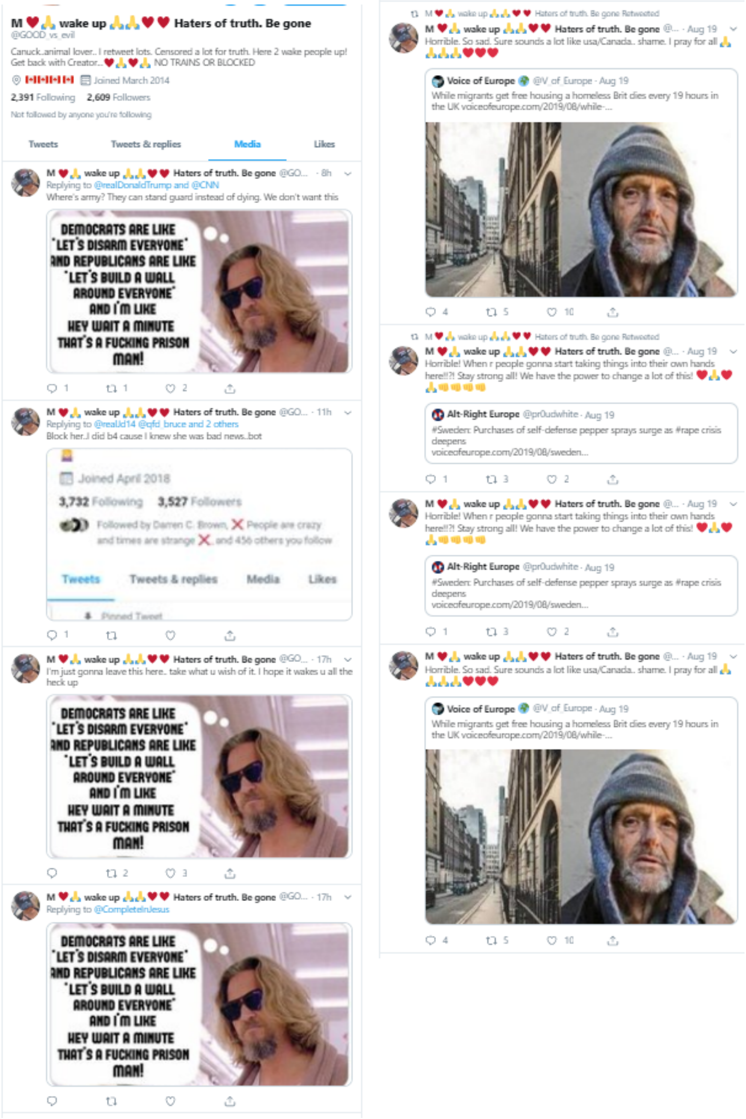
tweets has a duplicated SD5-hash. Examples of this in the Canadian election data is provided in Figure 6.8.

Popular Tweets - Unpopular Account

One technique used by a bot army is to have one or two accounts produce a tweet that is then 'pushed' by the bot army. Often the originator of the message isn't important, and can be selected at random. This creates the phenomena where you have an account that is not very popular producing tweets that appear very popular, which is counter intuitive. We call this the *pop-unpop* label, which is returned positive if any original tweets are 5 times more popular than the account (receives 5 times more retweets than the account has followers). This is illustrated in Figure 6.9.

Dormant User

Account dormancy is an important factor to take into consideration within information campaigns. Often bot and troll armies are activated for given events, and once done may remain dormant for long periods of time. In the case of `bot_labeler`, we create a positive label for account dormancy if the account has not produced any content (tweet, reply, retweet) in the last 6 months.



M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone
@GOOD_vs_evil

Canuck..animal lover.. I retweet lots. Censored a lot for truth. Here 2 wake people up! Get back with Creator...❤🙏❤🙏 NO TRAINS OR BLOCKED

🇨🇦🇨🇦🇨🇦🇨🇦 Joined March 2014

2,391 Following 2,609 Followers

Not followed by anyone you're following

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @GO... · 8h
Replying to @realDonaldTrump and @CNN
Where's army? They can stand guard instead of dying. We don't want this

1 1 2

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @GO... · 11h
Replying to @realJd14 @qfd_bruce and 2 others
Block her..I did b4 cause I knew she was bad news..bot

Joined April 2018
3,732 Following 3,527 Followers
Followed by Darren C. Brown, ❌ People are crazy and times are strange ❌, and 456 others you follow
Pinned Tweet

1

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @GO... · 17h
I'm just gonna leave this here.. take what u wish of it. I hope it wakes u all the heck up

2 3

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @GO... · 17h
Replying to @CompleteInJesus

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone Retweeted
M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @... · Aug 19
Horrible. So sad. Sure sounds a lot like usa/Canada.. shame. I pray for all 🙏🙏🙏🙏❤❤❤

Voice of Europe 🌍 @V_of_Europe · Aug 19
While migrants get free housing a homeless Brit dies every 19 hours in the UK voiceofeurope.com/2019/08/while-...

4 5 10

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone Retweeted
M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @... · Aug 19
Horrible! When r people gonna start taking things into their own hands here!!?! Stay strong all! We have the power to change a lot of this! ❤🙏❤🙏👊👊👊👊

Alt-Right Europe @pr0udwhite · Aug 19
#Sweden: Purchases of self-defense pepper sprays surge as #rape crisis deepens
voiceofeurope.com/2019/08/sweden...

1 3 2

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @... · Aug 19
Horrible! When r people gonna start taking things into their own hands here!!?! Stay strong all! We have the power to change a lot of this! ❤🙏❤🙏👊👊👊👊

Alt-Right Europe @pr0udwhite · Aug 19
#Sweden: Purchases of self-defense pepper sprays surge as #rape crisis deepens
voiceofeurope.com/2019/08/sweden...

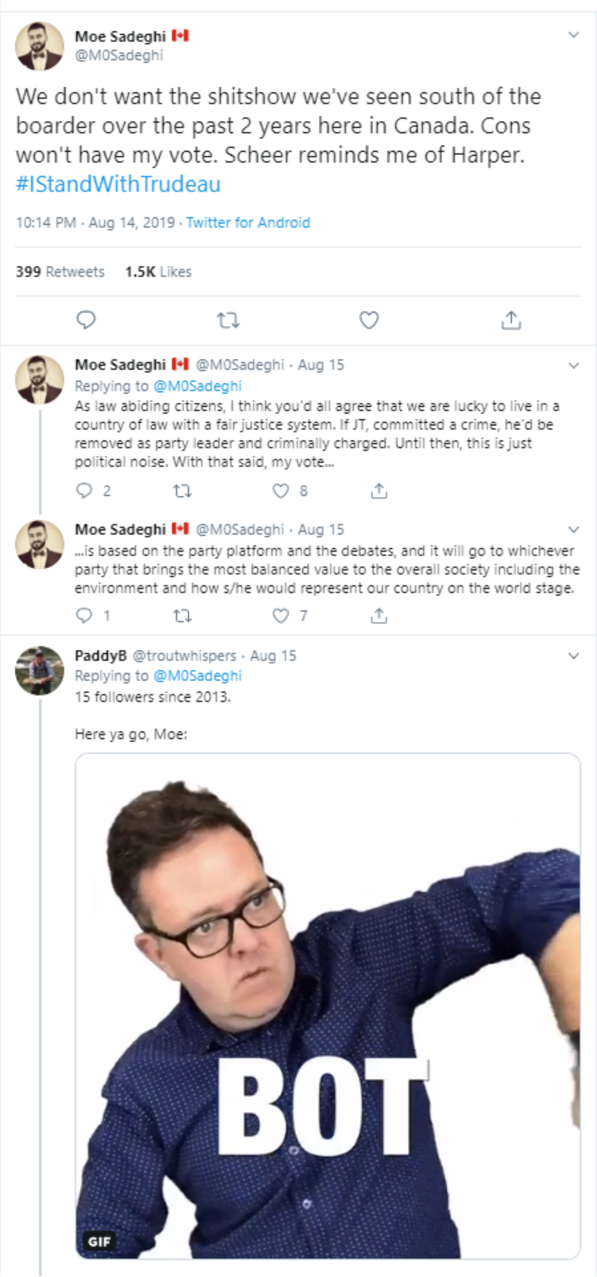
1 3 2

M ❤🙏 wake up 🙏🙏❤❤ Haters of truth. Be gone @... · Aug 19
Horrible. So sad. Sure sounds a lot like usa/Canada.. shame. I pray for all 🙏🙏🙏🙏❤❤❤

Voice of Europe 🌍 @V_of_Europe · Aug 19
While migrants get free housing a homeless Brit dies every 19 hours in the UK voiceofeurope.com/2019/08/while-...

4 5 10

Figure 6.8: Duplicate Pictures found on @GOOD_vs_evil bot account. Duplicate pictures illuminated this account that has produced 37K tweets meddling with both Canadian and US politics while amplifying Alt-right and conspiracy content



Moe Sadeghi 🇨🇦
@M0Sadeghi

We don't want the shitshow we've seen south of the boarder over the past 2 years here in Canada. Cons won't have my vote. Scheer reminds me of Harper. #IStandWithTrudeau

10:14 PM · Aug 14, 2019 · Twitter for Android

399 Retweets 1.5K Likes

Moe Sadeghi 🇨🇦 @M0Sadeghi · Aug 15
Replying to @M0Sadeghi
As law abiding citizens, I think you'd all agree that we are lucky to live in a country of law with a fair justice system. If JT, committed a crime, he'd be removed as party leader and criminally charged. Until then, this is just political noise. With that said, my vote...

2 8

Moe Sadeghi 🇨🇦 @M0Sadeghi · Aug 15
...is based on the party platform and the debates, and it will go to whichever party that brings the most balanced value to the overall society including the environment and how s/he would represent our country on the world stage.

1 7

PaddyB @troutwhispers · Aug 15
Replying to @M0Sadeghi
15 followers since 2013.

Here ya go, Moe:

Figure 6.9: This account is participating in the election conversation with pro-administration information, and only has 25 followers. This tweet has 399 retweets, and shows a thread where folks are identifying it as a bot because of its “unpop-pop” characteristics.

State-Sponsored Propaganda

Since recent articles and attention has been paid to state-sponsored use of bots in numerous events, particularly democratic events (Brexit vote, US Presidential and 2018 Midterm elections, Swedish elections, German elections, Hong Kong protests), we decided to apply a label if the account is propagating state-sponsored media. Examples of state-sponsored media include @RT_America and @SputnikInt from Russia, @XHNews from China, @IrnaEnglish from Iran, and @VOANews from America. While each handle is associated with a state-sponsored media entity, the role and level of independence of the entity varies widely. For example, Voice of America has a very different role than Sputnik. Similarly, organizations like Deustch Welle in Germany enjoy significantly more independence than XHNews in China.

In total, we collected 82 Twitter handles of known state-sponsored media. To apply this label, we search the content of the last 400 tweets by a user to see if they contain connections to state-sponsored media. If any state sponsored media was detected, we applied a label to indicate possible connection/amplification of state-sponsored propaganda. The metadata for this provide details of what nation(s) are being amplified. A Table of Tweets amplifying state-sponsored accounts in the Canadia 2019 National election is presented in Table 6.2. While roles and independence vary widely, we found this function helpful in identifying data with any connection to state-sponsored media, and the metadata allows the analyst to parse this further.

Table 6.2: Table of Total Tweets found in the Canada Data amplifying state-sponsored accounts.

Country	Example Handles	Tweets Amplifying State-Sponsored Accounts
China	@XHNews, @Xhdeutsch	3927
Germany	@dw_deutsch, @DeutscheWelle	720
Iran	@IrnaEnglish, @iribnewsFa	20
Israel	@NewsChannelIL	23
Korea	@KBSAmericaInc, @KBSWorldTV	19
Russia	@SputnickInt, @RTUKNews	4472
USA	@VOINews, @VOAAfrica	347

Questionable News Sources

The modern tsunami of disinformation has generated significant discussion and debate about *fake news*. This discussion has been politicized, and unfortunately has at times become synonymous with information operations. While we've found that false or fake news is only a small part of information operations, it is nonetheless present.

We wanted to have one of our labels indicate whether a given account is sharing questionable news and articles. We also wanted to be able to have meta-data that could indicate whether an account leans politically left (liberal) or right (conservative). To do this we collected and cleaned data provided by Media Bias Fact Check [3]. This website uses both crowd sourcing and trained analysts to provide analysis of various websites associated with news and/or science. This data

[3] https://mediabiasfactcheck.com

provides analysis of the bias of the website (left, left-center, center, right-center, right). It also indicates the level of factual reporting by the site (very high, high, mixed, low, very low). Finally, it indicates whether the site is associated with fake news, conspiracy theories, or satire. For our purposes, we mapped the factual reporting levels to numeric values: very high = 2, high = 1, mixed = 0, low = -1, and very low = -2.

To analyze a given Twitter Account, we extract URLs from the last 400 tweets shared by the account and then expand any URLs that have been shortened with the Bitly link shortening service. We then extract the domain name from the URL, and compare with the data from Media Bias Fact Check. The mean factual value of all found links is then computed. If the mean factual reporting value is lower than 0, the label is returned True for questionable news sources. The metadata returned with the label includes value counts for bias and for factual reporting labels. Bias and factual reporting tabulation for the 2019 Canadian national election dialogue on Twitter is provided in Table 6.3.

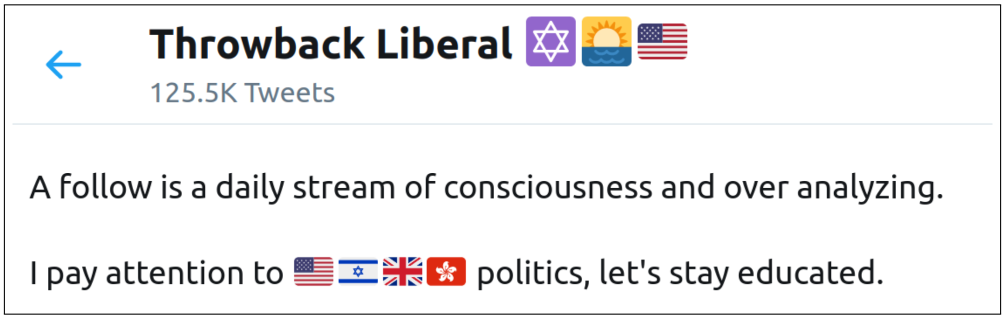
Table 6.3: Bias and Factual Reporting Tabulation for the 2019 Canadian National Election Twitter Dialogue.

		Bias					Other Labels			
		L	LC	C	RC	R	conspiracy	fake-news	satire	pro-science
Factual Reporting	very high	0	15K	5K	0	0	0	0	0	0.5K
	high	12K	247K	45K	102K	1K	0	0	0	3K
	mixed	9K	8K	0.1K	15K	92K	5K	0	0	0
	low	0	0	0	0	0	5K	0	29K	0
	very low	0	0	0	0	0	0.5K	28K	0	0

* L = *Left*, LC = *Left-Center*, C = *Center*, RC = *Right-Center*, R = *Right*

Multiple National Flags

In previous research we've highlighted how some malicious actors use multiple national flags in the account description of an account as if the account is an expatriate or frequent traveler [36]. This ruse is created to allow the account to manipulate the political process and online debate in multiple countries around the world without looking too out of place. An example of this phenomena in the Canadian elections is provided in Figure 6.10.



Throwback Liberal ✡️🌅🇺🇸

125.5K Tweets

A follow is a daily stream of consciousness and over analyzing.

I pay attention to 🇺🇸🇮🇱🇬🇧🇭🇰 politics, let's stay educated.

Figure 6.10: Example use of multiple flags by a liberal bot in the Canadian election. This allows it to more easily and seamlessly participate in multiple national conversations.

No Sleep

One of the classic attributes of automated accounts is non-human circadian rhythms. Bots don't sleep, and therefore they can render inorganic diurnal patterns. If an account lacks any circadian rythm, the distribution of its content by hour of day will slowly become a uniform distribution. To apply this label, we create an hour of day distribution with the last 400 tweets (the algorithm requires a minimum of 50 tweets). We then conduct the Kolmogorov-Smirnov non-parametric test for uniformity. A p-value greater than 0.5 provides strong evidence of non-human circadian rhythms [37]. The non-parametric test value is provided in the meta-data.

Abusive Language

Bots and especially trolls can produce or propagate abusive language. This language supports intimidation, aggravation, and other malicious intent. Inflammatory and abusive language provokes emotional responses and breeds discord in internet discussions. This discord can often overflow from the virtual to the real world.

We leverage a dictionary approach to finding abusive language. We clean and tokenize each tweet, and then lookup the language that Twitter attributes to the Tweet (this label is found in the raw Twitter JSON). We then lookup the abusive terms associated with this language from an abusive terms dictionary that merges efforts at Carnegie Mellon University and Netanomics Company with an open source dictionary shared on Github [4]. This dictionary contains 13,277 abusive terms from 63 languages. A positive abusive terms label is applied if more than 20% of 400 tweets contain abusive language. The user is provided with number of tweets that contain abusive language and the list of abusive terms found. These are provided in Figure 6.11.

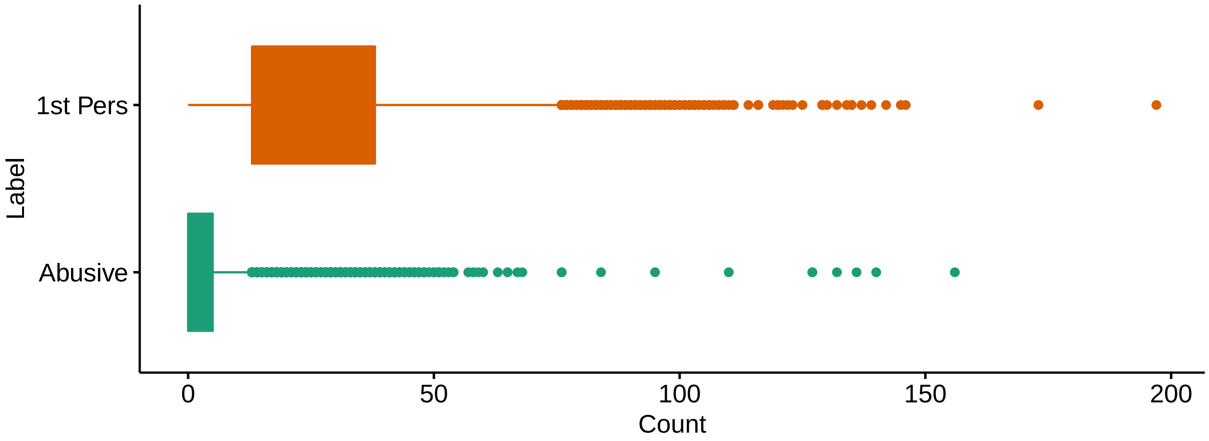


Figure 6.11: Count of Tweets containing abusive terms and 1st Person Pronouns

[4] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

1st Person Pronouns

1st person pronouns have been shown to reveal personality, maturity, emotional stability as well as other signals [191]. Higher use of personal pronouns demonstrates self-focus and are correlated with depression and lower status. Given this important signal, we measure personal pronouns used in the account's tweets.

Having cleaned and tokenized the text, we use a dictionary method in the same way that we did for abusive terms. In this case the personal pronoun dictionary was produced by a joint effort from Carnegie Mellon University and Netanomics Company. This list contains 914 personal pronouns from 63 languages. Unlike *Abusive Language*, however, this is only run on original tweets and replies that the account produces (it is not run on retweets). This returns true if more than 30% of the original tweets use a personal pronoun. Metadata includes the list of personal pronouns used. The tweet counts for Canada Data are provided in Figure 6.11.

6.5.3 Network Labels

The labels below focus on the user's network, particularly the followers. On Twitter an account can choose their "friends" (the accounts they follow), but not their followers. Who follows you reveals important information about an account.

Dormant Followers

Having applied our dormancy measure to accounts, we next apply it to the followers of the account. `Bot_labeler` returns the total number dormant accounts (no activity in last 6 months) for the last 5,000 accounts that followed the target account. If the percentage is higher than 25%, a positive label is applied. 25% was selected based on the percentage distributions provided in Figure 6.12.

Army of Silhouettes

Bots often attract bot followers, and therefore many of their follower pages have many accounts with default profile images, creating what is sometimes called an "Army of Silhouettes". We apply the 'Army of Silhouettes' label for any account that has more than 40% of their followers with a default profile image. Proportions of Silhouette bearing followers are provided in Figure 6.12.

Bot Followers

As indicated above, bots tend to attract other bots. In order to analyze this, we use the Bot-hunter supervised machine learning model [30] to determine whether the followers (up to the last 5,000) of the account have a high probability of being a bot. This model extracts features from the follower objects which are scraped from the REST API. The Bot-hunter model returns a probability of an account being a bot. We determined that a follower was a bot if the returned probability was over 0.6. We also determined that the target account has significant bot followers

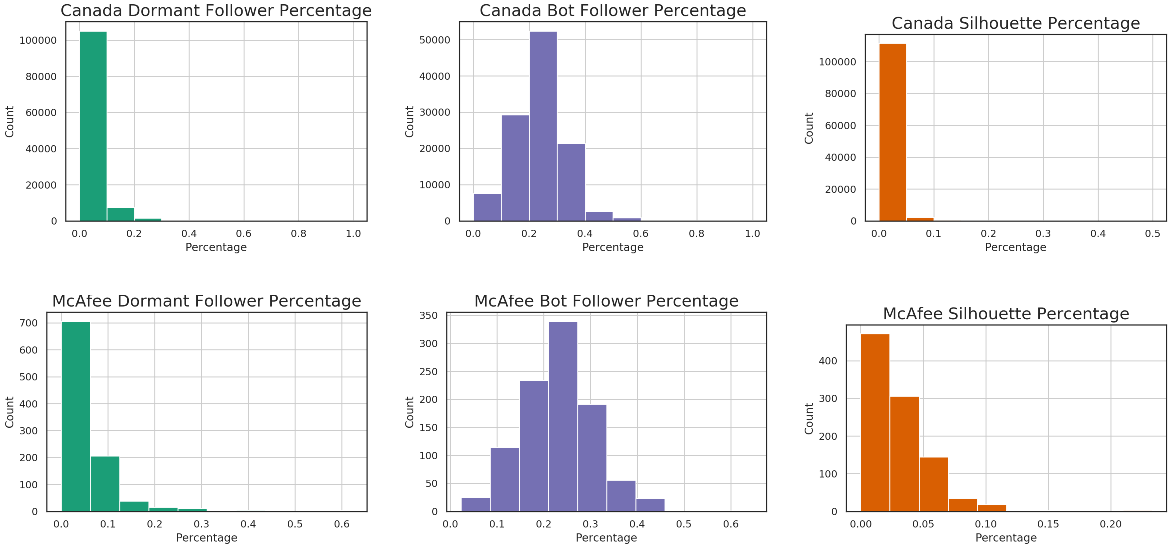


Figure 6.12: Distribution of Dormant Follower, Bot Follower, and Profile Picture Silhouette Percentages for Canada and McAFee Data

if over 20% of their followers were likely bots. Proportions of bot followers is provided in Figure 6.12.

6.5.4 Data Ouput

In addition to rendering and aggregating labels, `bot_labeler` writes the label metadata to disk in a JSON format. This allows an analyst to conduct deep dives, or if need be to change label thresholds where appropriate. An example JSON output is provided in Listing 6.1.

The `bot_labeler` Python package offers the ability to run only fast metrics (this primarily drops the image labels such as *high memes*). If the fast option was selected, the metadata will have a flag indicating this.

6.6 Results

Having run the `bot_labeler` methodology on both the Canadian and McAfee data, we provide a summary of the results in Table 6.4. This also acknowledge the threshold where appropriate.

In the Canada data we observe the 37% of bot accounts are amplifying other's messages and narrative. This seems to be a primary purpose of these bot accounts in the Canadian election discussion on Twitter. We also observe that 13% of them are sharing biased or content with a low factual value. We also see that 22% of these accounts are *high volume* (sharing more than 50 tweets per day). We also observe that 3.5% of the accounts are sharing or propagating state sponsored content and narratives. Overall this characterizes the bot involvement in the Canadian elections as a force multiplier attempting to amplify certain narratives with high volume bots, and contains limited overt state sponsored narratives.

```
{ 1st_pers_pronoun : { first_pronouns : [ yo ,  io ,  mi ,  nuestro ,  ours
   ],
                       label : False,
                       tweets_with_1st_pronoun : 10},
  abusive_terms : { abusive_terms : [],
                    label : False,
                    tweets_with_abusive : 0},
  amplification : { label : False,  retweet_percentage : 0.34},
  army_of_silhouette : { label : False,  num_fol : 93,  num_silhouette : 0},
  bias : { bias : { NaN : 92,  fake-news : 5,  left : 1,  leftcenter : 5},
           factual_reporting : { NaN : 92,
                                 high : 5,
                                 mixed : 1,
                                 very low : 5},
           factual_value : { -2.0 : 5,  0.0 : 1,  1.0 : 5,  NaN : 92},
           label : True},
  bot_followers : { label : False,  num_bots : 17,  num_fols : 93},
  default_profile : { label : False},
  dormant_followers : { dormant_perc : 0.010752688172043012,  label : False
   },
  dormant_user : { label : False,
                   last_tweet_date :  Tue Nov 26 02:19:44 +0000 2019 },
  fast : True,
  flags : { flags : [],  label : False},
  gender_mismatch : { age : [44.102740939901196],
                      label : False,
                      name_gender :  male ,
                      pic_gender : [{ gender :  male ,
                                      score : 0.016841834411025047}]},
  high_vol : { label : False,  mean_daily_tweets : 24},
  honeypot : { label : False,  nsfw_score : 0.0006152272690087557},
  no_sleep : { KS-stat : 1.5810657226861515e-15,  label : False},
  propaganda : { label : True,  propaganda : { Russia : 21}},
  random_lang : { label : False,
                  language_dictionary : { cy : 1,
                                          en : 8,
                                          es : 170,
                                          it : 1,
                                          pt : 2,
                                          und : 18}},
  spam : { label : False,  median_mentions : 1.0},
  unpop-pop : { followers : 44261,  label : False,  max_retweets : 3299}}
```

6.1: Example bot_labeler Output

Table 6.4: Bot Labeler Results for McAfee and Canada Data

Label	Threshold	McAfee	Canada
1st_pers_pronoun	$>$ 30% original tweets contain 1PP	9.5%	6.0%
abusive_terms	$>$ 20% contain abusive words	0.0%	0.3%
amplification	More than 70% retweets	4.2%	37.4%
army_of_silhouette	40% Default Profiles in Followers	0.0%	0.0%
bias	Mean Factual Value < 0	7.1%	13.0%
bot_followers	More than 20% bot followers	66.8%	66.3%
default_profile	If true	0.8%	11.3%
dormant_followers	More than 50% dormant	0.3%	0.1%
dormant_user	Is dormant	0.0%	0.1%
flags	2 or more national flags	0.0%	1.0%
gender_mismatch	Mismatch	1.9%	1.9%
high_vol	More than 50/day	1.2%	22.9%
honeypot	0.3 NSFW Probability	37.4%	1.3%
no_sleep	If KS-statistic > 0.5	2.2%	0.0%
propaganda	Any State Sponsored	0.2%	3.5%
random_lang	> 5 with more than 1	1.2%	2.5%
spam	median 3 mentions	0.0%	0.7%
unpop-pop	Tweets 5x more popular than account	0.0%	0.8%
duplicate_pictures	Any Found	1.2%	**
high_memes	20% of images	3.9%	**

** Not available in *fast* version of bot_labels.

Since these are non-exclusive labels, we see several overlaps in the data. For example, in the Canada data we see high amplification being conducted by high volume accounts that have a high bot presence and bot followers. In the McAfee bot net we see honeypot accounts with a high bot followers that are propagating the bitcoin scam. We explore the correlation between labels in the correlation heatmap provided in Figure 6.13. Here we see positive correlation between dormant users and default profile as well as amplification accounts and high volume accounts. We also see negative correlation between high volume accounts and default profile as well as other pairs.

In the McAfee data we observe that 37% of these accounts are labeled as *honey pot* accounts, and are using risqué images to draw attention and drive actions. We also see that 67% of the the followers of these accounts exhibit bot-like characteristics, highlighting the social manipulation present in this online scam.

6.7 Conclusion

Social cyber security workflows require a general and repeatable method to characterize malicious or otherwise anomalous accounts. Bot_labeler provides a method to apply non-exclusive labels to accounts to understand the general nature of their profile, content, and connected network. This can characterize either a handful of accounts or a larger batch of malicious accounts

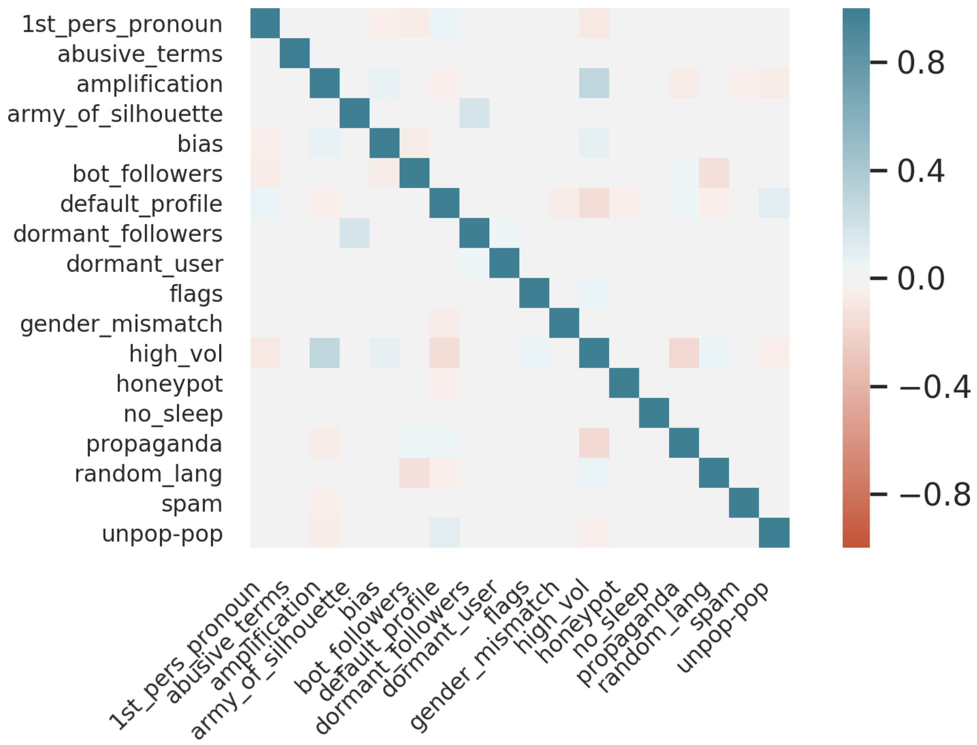


Figure 6.13: Correlation between labels for Canada Data

discovered with a traditional bot classification algorithm.

The *character* and *attributes* of malicious accounts gives insight into their creator and the purpose and intended target audience for which they were created. Combined this helps social cyber security practitioners understand and defend against this emerging threat. Sun Tzu's ancient military advice claims that ”If you know the enemy and know yourself, you need not fear the result of a hundred battles.” Bot_labeler allows leaders to understand the enemy in information warfare. Now all we must do is know ourselves.

Chapter 7

Characterizing Campaigns (Sketch-IO)

Social cybersecurity analysts require the capability to rapidly delineate emerging information warfare campaigns. This includes mapping the effort across multiple social media platforms, identifying prominent themes, delineating targeted networks and populations, and measuring success. This type of campaign analysis needs to combine and enrich "tactical" tools such as bot, troll, and meme detection. In this chapter we present Sketch-IO, a proposed framework and supporting computational tools to automatically analyze streams that contain social media-based disinformation campaigns.

7.1 Introducing Campaign Analysis

Information warfare campaigns are right at our fingertips and yet difficult to grasp. As a myriad of actors participate in online conversations, it creates an ocean of data that is a mixture of various social media streams and conversations. In this ocean of data, it is non-trivial to separate steams and campaigns. In this chapter we will discuss pulling all of our research together to delineate and map information warfare campaigns.

In 2016 United States Department of Defense indicated their top information warfare related effort was to "*improve the capability of the Department to monitor, analyze, characterize, assess, forecast, and visualize the [Information Environment]*" [185]. This remains a top priority among many nations as they place increasing importance on defending their society and way of life from external manipulation.

There are several perspectives that we could use to map information warfare campaigns. In this chapter we will describe one perspective and use it to create a framework and develop methods to automate information campaign analysis. At the end of the chapter we will discuss other perspectives that could be used in a similar approach.

This chapter will start by discussing related military doctrine and research. It will then present and detail the Sketch-IO framework and automatic analysis tool that our team developed. This will include describing a prototype dashboard that we developed and tested. Finally, we present other information warfare frameworks that could be used to gain different views of an IO campaign.

7.2 Related Work

The most recent Army training manual "The Conduct of Information Operations" [121] defines the information environment, highlights interaction between the physical environment, information environment, and cognitive decision making, and details the information organization structure and operations process. This document is primarily oriented at the tactical level of war, and does not address characteristics of modern information campaigns that are emerging in the online information environment (OIE). Joint Publication 3-13 "Information Operations" goes a bit farther in using this framework to define target audiences, but does not go into detail on how Information Related Capabilities (IRC) are applied to accomplish detailed ends [222]. JP 3-13 uses the Graphic in Figure 7.1 to describe the environment, trying to relate the virtual environment, human/social environment, and physical environment as well as place all actors into three categories: *key influencers*, *vulnerable populations*, and *mass audiences*. While helpful in understanding the environment, this framework is too general and does not describe the "information flow" between actors.

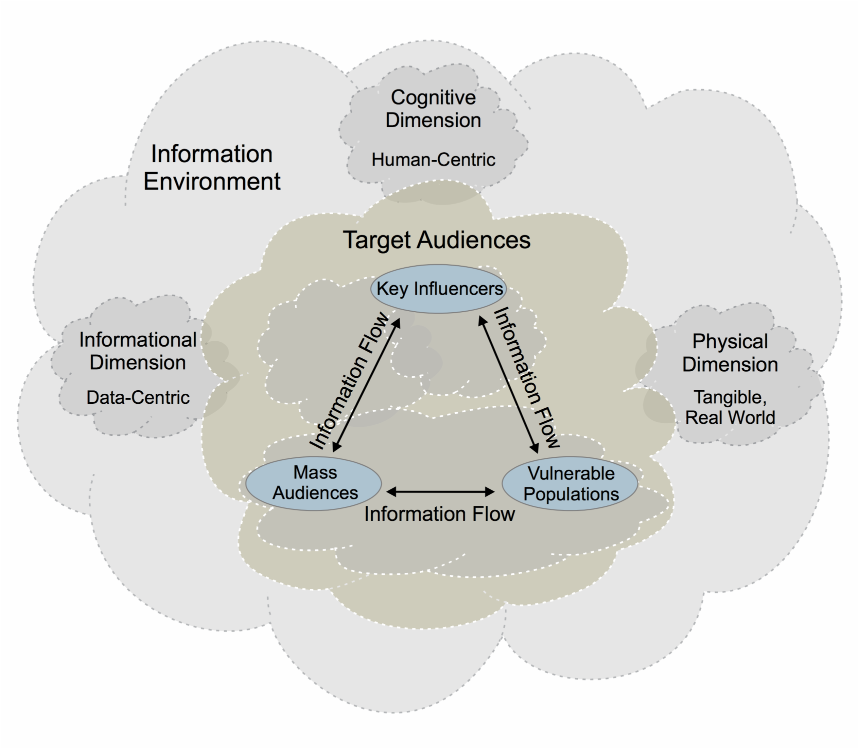


Figure 7.1: Describing information environment and target audiences in Joint Publication 3-13 [222]

.

Waltz provides a thorough systems engineering approach of information warfare in his 1998 book "Information warfare: Principles and operations" [244]. He describes information warfare principles from both government and academic disciplines. In regard to frameworks, he defines information warfare principles by combining basic data science pipelines with the Observe, Orient, Decide, Act (OODA) loop originally proposed by Boyd [50]. Walz also combines strategic processes with assessment (see Figure 7.2) that we intend to partially replicate in the modern information environment. Waltz does not address social media (it did not exist in 1998), and he is also focused on military against military, without fully appreciation the use of information warfare against a given population or larger society.

Alberts focuses on defensive information warfare [12], but also falls short of providing a framework for the age of social media. From marketing literature, Kim identifies a four step model for social media marketing: 1) Listening 2) Strategic Design, 3) Implementation and Monitoring and 4) Evaluation [139] from which we will derive some of our basic elements for Sketch-IO.

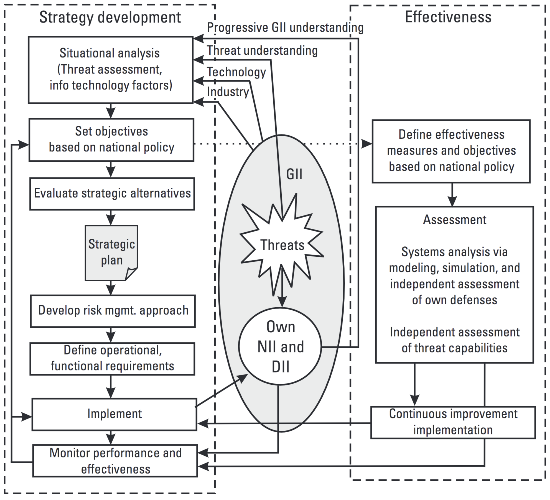

Strategy development
Effectiveness
Progressive GII understanding
Threat understanding
Technology
Industry
Situational analysis (Threat assessment, info technology factors)
Set objectives based on national policy
Evaluate strategic alternatives
Strategic plan
Develop risk mgmt. approach
Define operational, functional requirements
Implement
Monitor performance and effectiveness
GII
Threats
Own NII and DII
Define effectiveness measures and objectives based on national policy
Assessment
Systems analysis via modeling, simulation, and independent assessment of own defenses
Independent assessment of threat capabilities
Continuous improvement implementation

Figure 5.2 The strategic process includes strategy development and assessment elements.

Figure 7.2: Waltz uses this graphic to describe the strategic development, implementation, and evaluation in information operations (This is from [244]). He defines GII as Global Information Infrastructure, NII in National Information Infrastructure, and DII is Defense Information Infrastructure.

Many of the works discussed above do not take into consideration the nature of online social media. Today's online environment does not have clear boundaries between civil and military. Civil, threat, and friendly actors are all mixed, and at times find themselves supporting similar narratives and voices (though for different reasons). This type of virtual battlefield creates challenges for military leaders who prefer clear boundaries and associated authorities.

7.3 Delineating Information Campaigns

The offensive information warfare campaign coordinator as well as the corporate social media marketing manager have a clear list of all their accounts and messages, and can easily identify what part of the larger conversation is theirs. For the defensive information warfare analyst, however, extracting a coordinated malicious campaign from a larger conversation is not trivial. Information warfare campaigns are designed to naturally blend in with the conversation, and often come along side an existing social, political, or alternative movement. Trying to separate malicious actors and messages from these existing grassroots movements is difficult.

Several methods for filtering social media are available and can be used by themselves or in combination with other methods. These are listed and described below:

1. Key word filtering [147]
2. Snowball sampling [111]
3. Community detection [28]
4. Bot-match and graph/content embedding (Chapter 4)
5. Stance prediction [146]

Key word filtering involves using unique semantic tokens or a combination of tokens to extract a specific information campaign. A coordinated campaign will use unique hashtags, URLs, screen names, or other semantic tokens that can be used by themselves of in combination to identify a specific campaign. This key word filtering is often incorporated into the initial collection method, such as using these keywords on Twitter Streaming or REST API. Analysts can also dynamically update these terms as illustrated by Kumar [147]. Snowball sampling, originally proposed by Goodman in 1961 [111], can also be used to extract the network associated with an information campaign. Snowball sampling is used when an analyst can only identify one or more malicious accounts associated with a campaign. In snowball sampling, these malicious accounts are used as seed nodes, and the analyst then finds links to other accounts, and continues to recursively "snowball" to additional links until he/she is sufficiently satisfied that they have found the network of interest. Keyword filtering and snowball sampling are the primary collection methods to get an initial data stream or network.

Once the analyst has a stream or network, they can continue to extract the core campaign using community detection, embedding techniques, and stance prediction. If the underlying social network ($links = friends|following$ connection) or communication network ($links = retweet|reply|mention$) has clusters that are predominantly malicious actors, then community detection can be used to extract these clusters. While numerous community detection algorithms exist, our team has found that the louvain algorithm [46] is effective and scales to large data. If trying to extract dense subgraphs associated with echo chambers, we also leverage a dense subgraph detection algorithm by Benigni [27]. Various embedding techniques can be used to find "similar" accounts given a few seed accounts. These embedding techniques are described in detail in Chapter 4. Stance prediction as proposed by Kumar [146] can be used if the information campaign is aligned with a given stance in a polarized debate. This may be a stance on a controversial topic (abortion, climate change, etc) or may be a stance on a given conspiracy theory. Stance prediction can be used to extract all users who align with a given stance on a topic.

Together the above techniques are used to extract and reduce the data so that a given data set is largely associated with a given information campaign. This delineated campaign data becomes the input for Sketch-IO. The value derived from Sketch-IO relies heavily on the accuracy of the above methods in delineating the campaign. Data that is poorly delineated introduces significant noise into all analytics used for Sketch-IO.

7.4 Sketch-IO Framework

In Chapter 2 we detailed the various forms of maneuver in the narrative and network. The goal of Sketch IO is to map this manipulation to the information environment. Where is the content (text, image, video) being created and posted? How is this mapped to a network in mainstream social media (Twitter, Facebook, etc)? How is this supported with funding and organization? And how successful is the campaign? The template for this is provided in Figure 7.3. Below we will discuss each element of this framework. Later we will discuss operationalizing this with specific computational tools against a given information campaign.

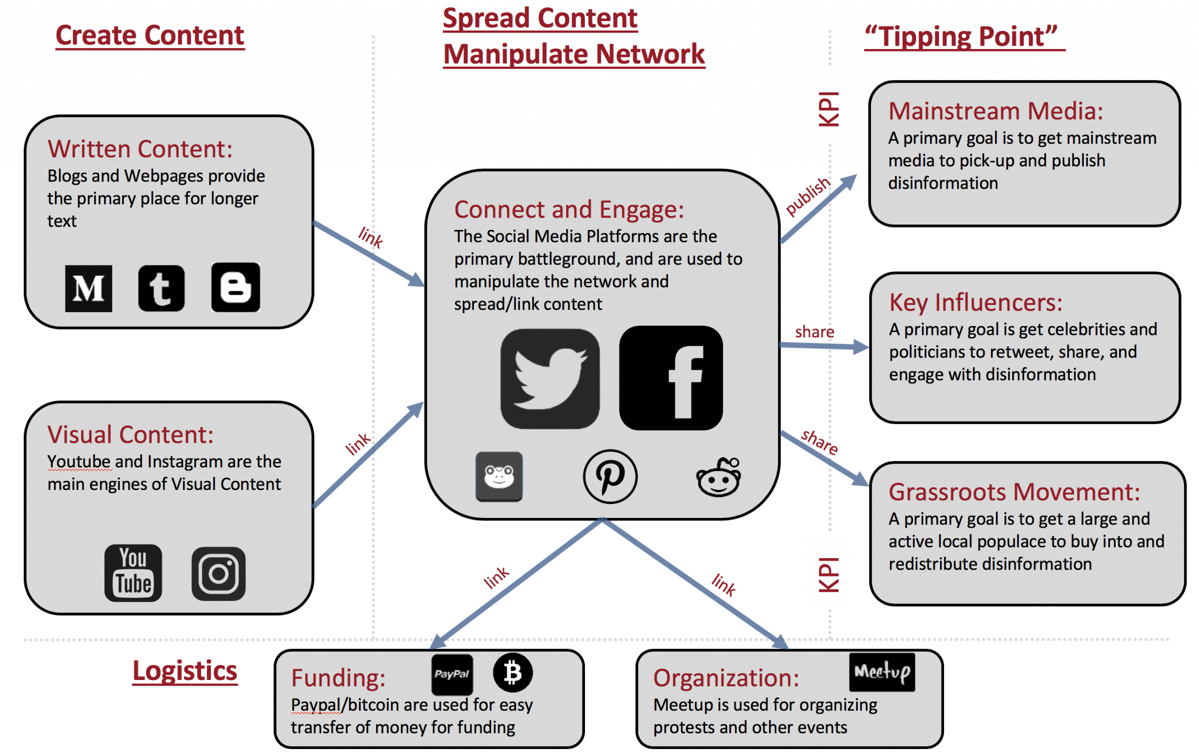

Create Content
Spread Content
Manipulate Network
"Tipping Point"
Written Content:
Blogs and Webpages provide the primary place for longer text
Visual Content:
Youtube and Instagram are the main engines of Visual Content
Connect and Engage:
The Social Media Platforms are the primary battleground, and are used to manipulate the network and spread/link content
Mainstream Media:
A primary goal is to get mainstream media to pick-up and publish disinformation
Key Influencers:
A primary goal is get celebrities and politicians to retweet, share, and engage with disinformation
Grassroots Movement:
A primary goal is to get a large and active local populace to buy into and redistribute disinformation
link
link
publish
share
share
link
link
KPI
KPI
Logistics
Funding:
Paypal/bitcoin are used for easy transfer of money for funding
Organization:
Meetup is used for organizing protests and other events

Figure 7.3: The Sketch-IO Framework

7.4.1 Create Content/Narrative

One of the key tasks in mapping a known and delineated information campaign is identifying the source and popularity of multimedia content. Where are actors creating text, image, and video content in order to connect it with the target audience? Text content includes blogs (Medium, Blogger, WordPress, etc), traditional web sites, micro-blogs (Twitter, Gab, Sino Weibo), and messaging apps (Facebook Messenger, WhatsApp, Telegram). This text may be created by the IO organization, or it may be created by another actor and simply "pushed" by the IO organization.

Any online marketing or information campaign requires engaging visual content in order to attract attention. Increasing efforts are put into creative, engaging, attention-grabbing and at times disturbing multi-media. Consumers often attribute increased veracity with increased quality of the video, and so those who sow disinformation attempt to create professional videos to get their deceit to stick.

While multi-media (images and video) can be uploaded directly to a specific social media platform (Twitter, Facebook, etc), it is often more efficient to load it to a site such as Instagram or Youtube and then link it directly to multiple platforms. In addition to uploading once and only having a single repository to update, the originator can collect engagement metrics for a single piece of multimedia from across platforms.

7.4.2 Spread Content & Manipulate the Network

Social media platforms are the primary battleground where actors infiltrate and manipulate networks in order to prepare them for their narrative. For manipulation in the West, Facebook and Twitter reign supreme, with significant effort leveraged against both platforms. Efforts also exist on Gab, Pinterest, Reddit, and others, though at a lesser extent. In some parts of the world (Africa/Asia), WhatsApp has become a critical attack vector for disinformation. It is on this battlefield that actors conduct the BEND forms of maneuver outlined in Chapter 2. It is here that they *build*, *bridge*, *nuke* and *narrow* networks to prepare them for a specific narrative. This is where bots serve as force multipliers to push certain narratives and cyborgs and trolls incessantly attack other narratives and accounts.

In understanding the battlefield, we attempt to understand how the pieces and parts identified as content are mapped into networks and conversations. What are the target communities and topic groups? How are overt actors such as state media participating in the operation?

7.4.3 Key Performance Indicators: the “Tipping Point”

For information operations, determining whether or not you’re successful is difficult. Although often compared to a marketing campaign, information operations lack hundreds of thousands of sales events that provide detailed metrics of performance. Nike, for example, can control most other variables, then introduce a marketing campaign, and confidently attribute changes in the revenue stream to the marketing campaign. The marketing campaign has changed the beliefs and behavior of people, and in this case they vote with their wallet. Information campaigns don’t have a similar method to extract the true beliefs of an individual by forcing them to “put skin in the game.” Polls are relatively infrequent and don’t require participants to put any “skin in the game.” On top of this, the evolution of collective beliefs and norms in a large culture are extremely complex and the result of a myriad of variables that are difficult to control for. There are, however, emerging behaviors that are indicators of success.

Performance indicators are defined “...as an item of information collected at regular intervals to track the performance of a system” [96]. From our study of information campaigns, it appears that IO organizations focus on a couple key performance metrics (KPI’s) by which they track performance. The goal is to get their message or narrative to begin moving on its own. Evidence of this is 1) the narrative is picked up by mainstream news or news aggregators, 2) the narrative is propagated by key influencers (politicians, celebrities, etc), or 3) the narrative is picked up and becomes a key tenet in a new or existing grassroots movement. We will discuss each of these below.

One of the key measures of success for an information campaign is to get mainstream media, media aggregators, or even alternative media to promote the narrative or story. Mainstream media includes primary media companies (ABC, NBC, Fox, New York Times, BBC, etc). News aggregators include services like Feedly, Google News, Flipboard, ThinkProgress, and the Drudge Report. These can be extremely influential, as evidenced by the fact that the Drudge Report jumped ahead of the New York Times in online popularity in 2018 [47]. Alternative News includes left leaning sites like Democracy Now! and the Jacobin and right leaning sites like Breitbart and the Sean Hannity Show. In some cases an information campaign may prefer the alternative media outlets since stories there may receive less scrutiny inside a left leaning or right leaning echo chamber. An information campaign narrative gains instant credibility and broad reach when it is picked up by these news outlets or news aggregators.

The next key measure of success is for an information campaign to receive direct support (think retweet) from key influencers such as celebrities, politicians, bloggers, or other online influencers. In the 2016 election campaign, Russian trolls were retweeted by 40 different politicians and celebrities, to include Presidential nominee Donald Trump and Twitter CEO Jack Dorsey [194]. This once again gives these actors and their narratives instant credibility and broad reach.

Finally, information warfare organizers attempt to launch a grass roots movement both virtually and in the physical world. This can often occur by tying their narrative to existing grass roots movements or subcultures (liberal groups, conservative groups, religious groups, LGBTQ groups, etc). These groups already have strong organization both online and offline, and it is easier to piggyback on their organization than to try to launch an entirely new movement.

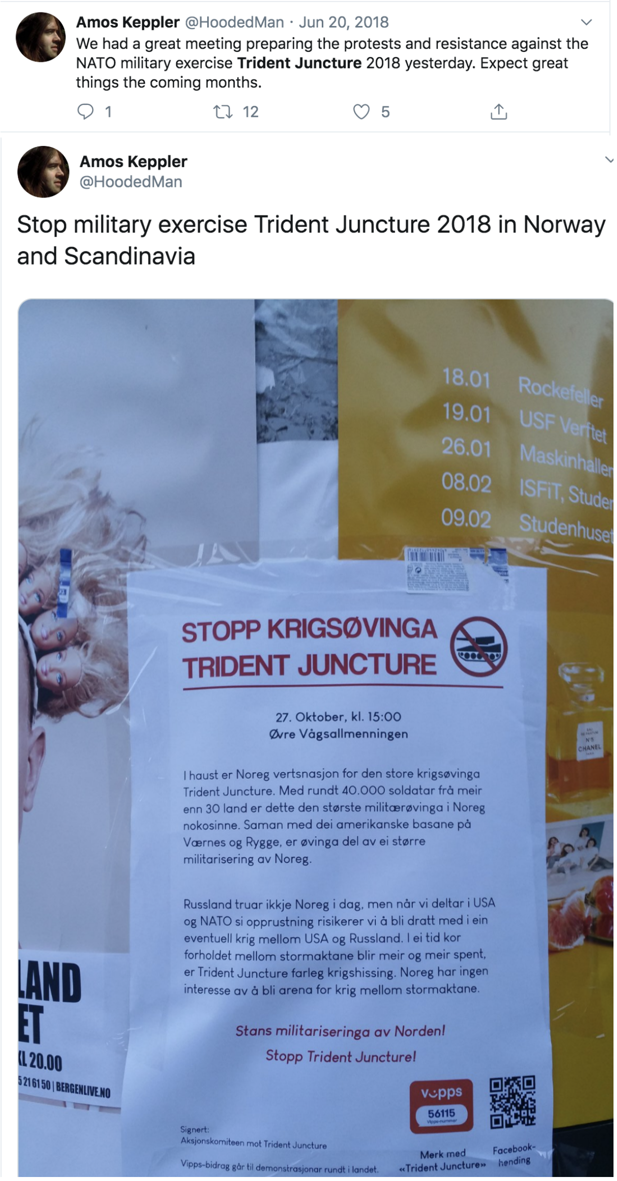
7.4.4 Logistics and Organization

Information operations actors may try to financially support their effort through online solicitation. Financial support can be solicited using Paypal, GoFundMe, and Bitcoin technologies. These solicitations are often staffed through the social media platforms to possible donors.

The overarching goal of information operations is for the virtual world to have an effect on the physical world. Sometimes this is more direct, such as when the information operations attempts to organize physical events (protests, meetings) through these platforms. In these instances the information campaign is attempting to directly coordinate and start a grass roots movement. At times this coordination is accomplished through platforms like MeetUp, but at other times it is coordinated directly through the social media platforms. In Figure 7.4 we see examples of a bot coordinating anti-NATO protests in Scandinavia.

7.5 Operationalizing Sketch-IO

In addition to developing a helpful mental framework for social cybersecurity professionals to use to envision information campaigns, our goal is to operationalize the framework by filling



Amos Keppler @HoodedMan · Jun 20, 2018

We had a great meeting preparing the protests and resistance against the NATO military exercise **Trident Juncture** 2018 yesterday. Expect great things the coming months.

1 12 5

Amos Keppler
@HoodedMan

Stop military exercise Trident Juncture 2018 in Norway and Scandinavia

STOPP KRIGSØVINGA TRIDENT JUNCTURE

27. Oktober, kl. 15:00
Øvre Vågsallmenningen

I haust er Noreg vertsnasjon for den store krigsøvinga Trident Juncture. Med rundt 40.000 soldatar frå meir enn 30 land er dette den største militærøvinga i Noreg nokosinne. Saman med dei amerikanske basane på Værnes og Rygge, er øvinga del av ei større militarisering av Noreg.

Russland truar ikkje Noreg i dag, men når vi deltar i USA og NATO si opprustning risikerer vi å bli dratt med i ein eventuell krig mellom USA og Russland. I ei tid kor forholdet mellom stormaktane blir meir og meir spent, er Trident Juncture farleg krigshissing. Noreg har ingen interesse av å bli arena for krig mellom stormaktane.

Stans militariseringa av Norden!
Stopp Trident Juncture!

Signert:
Aksjonskomiteen mot Trident Juncture

Vipps-bidrag går til demonstrasjonar rundt i landet.

vipps 56115 Vipps-nummer

Merk med «Trident Juncture»

Facebook-hending

Figure 7.4: A bot/cyborg account organizes and promotes protests against a planned NATO exercise in Scandinavia.

in each of the framework components with details from a given information campaign. This is illustrated in Figure 7.5.

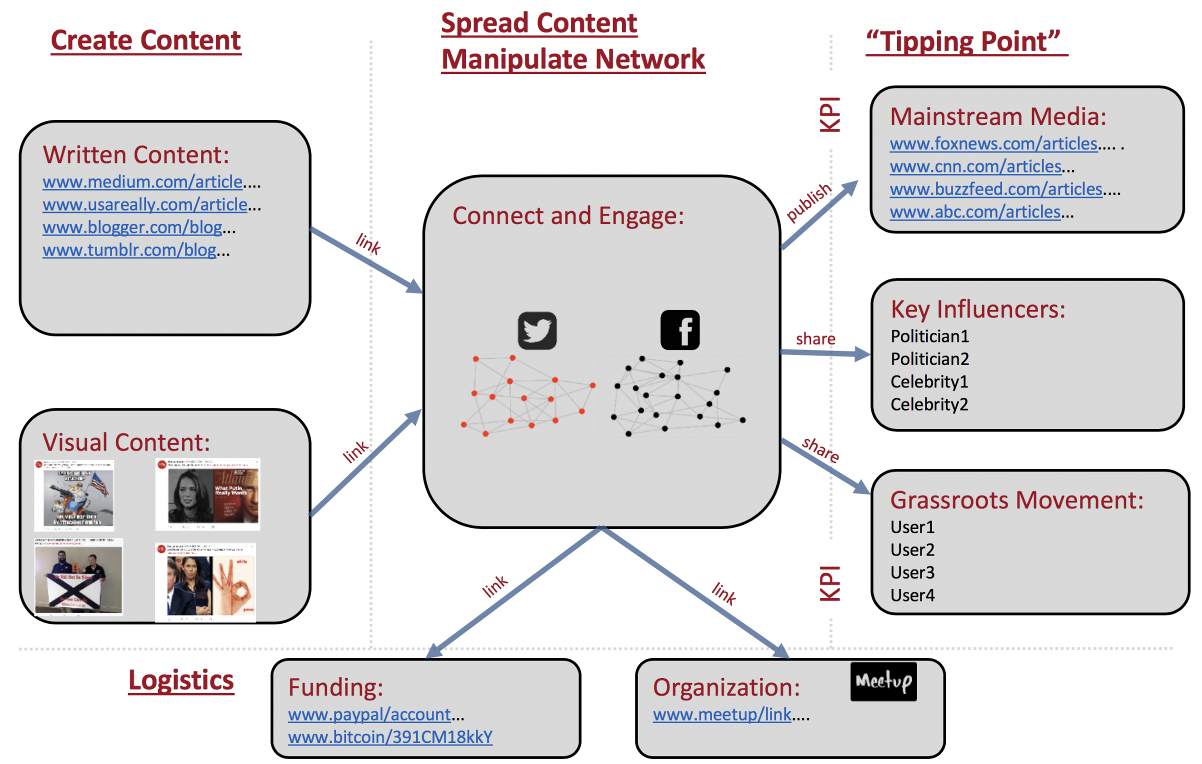


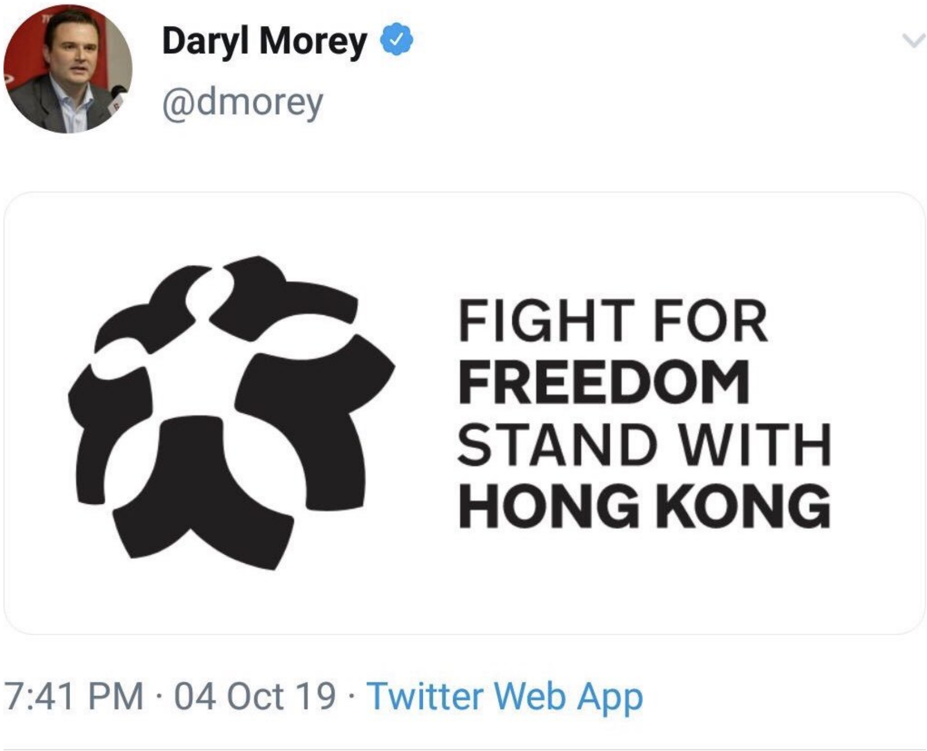
Figure 7.5: To operationalize the Sketch-IO framework, we begin to populate the framework as illustrated here

In this section we will discuss how we operationalized the prototype dashboard and tested the dashboard on the Chinese disinformation campaign focused on Hong Kong and the National Basketball Association (NBA). This information operation event is discussed in more detail below.

7.5.1 Hong Kong, the NBA, and Chinese Information Operations

Originally a British Colony, Hong Kong came under Chinese authority in 1997. Hong Kong is governed with a more lenient system that is often called the "one country, two systems" policy. In June 2019 Hong Kong residents began protesting several proposals that would allow extradition to mainland China. These protests grew in scope and persisted, eventually becoming violent with police using lethal ammunition and protesters using homemade "petrol bombs." The Chinese Government has previously used social media information operations to support their narrative and objectives in Hong Kong. In two separate efforts Twitter has suspended and released data associated with these information operations [203].

On 4 October 2019 Daryl Morey, the General Manager of the Houston Rockets, posted the Tweet seen in Figure 7.6 declaring his support for the Hong Kong protesters. This resulted in swift backlash from mainland China, which severed many business relationships with the Houston Rockets and the National Basketball Association (NBA) [253]. Morey and the NBA were targeted by a deliberate information campaign perpetrated by bots and trolls on social media [66]. Since many of these bots and trolls used Mr. Morey's screen name, we were able to use this as a search term to filter much of this disinformation campaign on Twitter.



Daryl Morey
@dmorey

FIGHT FOR
FREEDOM
STAND WITH
HONG KONG

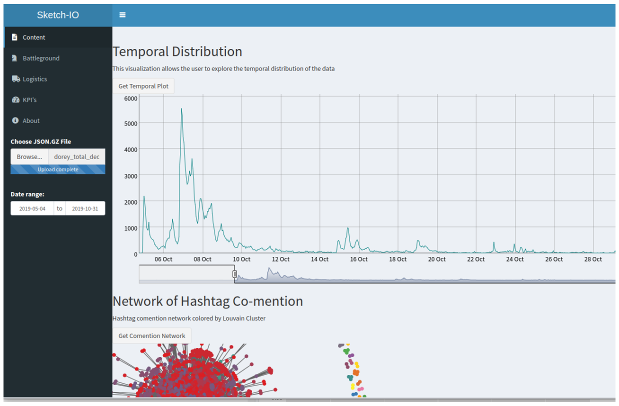
7:41 PM · 04 Oct 19 · Twitter Web App

Figure 7.6: The Tweet from Daryl Morey, General Manager of the Houston Rockets that incited a Chinese information operation

This gave us a delineated stream containing a deliberate information operation largely perpetrated by mainland China (though other actors could have capitalized on the event).

7.5.2 Identifying Engaging Content Across Platforms

The first step in characterizing an online social media campaign with the Sketch-IO framework is to identify the source content that is being mapped to a target audience. This includes the multi-media that is found across multiple platforms. Since these are being mapped into a target audience in the big social media platforms (Twitter, Facebook, etc), we can use the URLs in these streams to back into the original hosting sites and their respective media/content. In our prototype Sketch-IO app, we identify trending Youtube links, trending Facebook pages, and trending URL and domain names. This can be expanded to other trending sites of interest. Demonstrations of this can be found in Figure 7.7 and Figure 7.8. The content exploration tab also provides a temporal density plot as well as a dynamic network plot of the hashtag co-mention network. The tables allow full exploration of the URLs, allowing the user to provide custom search queries. Additionally, the user can expand Bitly shortened links (by default these are not expanded since it can be computationally expensive to expand the links).



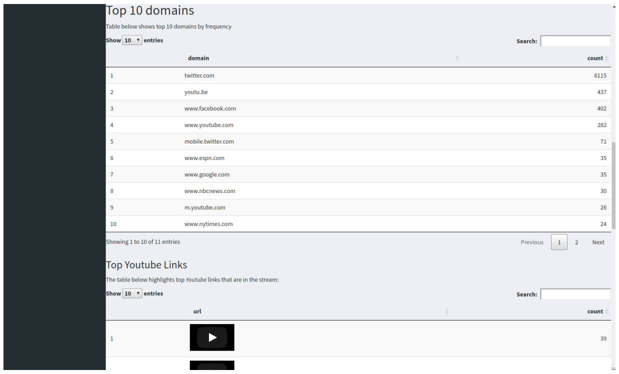
Temporal Distribution

This visualization allows the user to explore the temporal distribution of the data

Network of Hashtag Co-mention

Hashtag comention network colored by Louvain Cluster

Figure 7.7: The Initial Landing page maps temporal patterns and hashtag comention plots



Top 10 domains

Table below shows top 10 domains by frequency

Show 10 entries Search:

	domain	count
1	twitter.com	6115
2	youtu.be	437
3	www.facebook.com	402
4	www.youtube.com	282
5	mobile.twitter.com	71
6	www.espn.com	35
7	www.google.com	35
8	www.nbcnews.com	30
9	m.youtube.com	26
10	www.nytimes.com	24

Showing 1 to 10 of 11 entries

Top Youtube Links

The table below highlights top *Youtube* links that are in the stream:

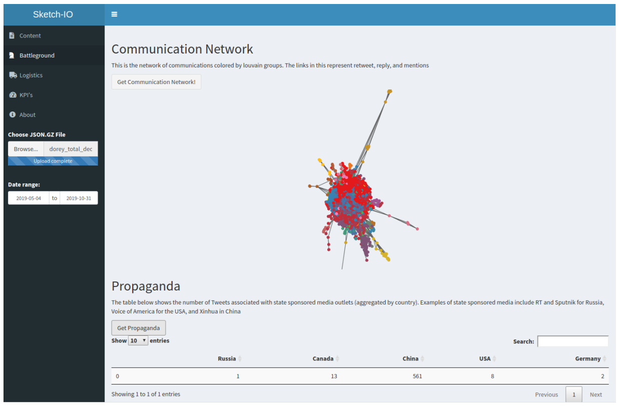
Show 10 entries Search:

	url	count
1	▶	39

Figure 7.8: The content analysis identifies trending material from other platforms such as Youtube and Facebook

7.5.3 Identifying Methods of Network Manipulation and Narrative Dispersal

Having identified cross platform content that is trending in the campaign, the next step is to map content into virtual communities and topic groups. The “Battlefield” tab on the Sketch-IO prototype allows this exploration. It starts by providing the user with a network visualization of the largest component in the communication network. A communication network is defined as a network based on directed communication links in social media. In Twitter this includes directed links associated with *retweets*, *mentions*, and *replies*. This network is colored by community as detected by the Louvain community detection algorithm [46]. This interactive visualization allows the user to determine the macro shape and density of the conversation as well as identify central and fringe communities in the conversation. A screen capture of this part of the Sketch-IO prototype is provided in Figure 7.9.



Communication Network

This is the network of communications colored by louvain groups. The links in this represent retweet, reply, and mentions

Propaganda

The table below shows the number of Tweets associated with state sponsored media outlets (aggregated by country). Examples of state sponsored media include RT and Sputnik for Russia, Voice of America for the USA, and Xinhua in China

	Russia	Canada	China	USA	Germany
0	1	13	561	8	2

Showing 1 to 1 of 1 entries

Figure 7.9: The “Battlefield” tab outlines the manipulation of the network and methods to connect the narrative with that network. It also includes community analysis and topic group analysis.

The next analytic provided is an algorithm that determines if state sponsored media outlets are involved in the conversation. Involved means they are either authoring posts or are retweeted, mentioned, or replied to in the conversation. The results are aggregated by country. Examples of state sponsored media include RT and Sputnik from Russia, Voice of America from the United States, and Xinhua in China. While each of these vary widely in purpose, intended audience, level of independence, and method of dispersion, they nonetheless provide valuable insight. For example, in an analysis of the China NBA disinformation effort, it is interesting that China state-sponsored media were found over 500 times in the data.

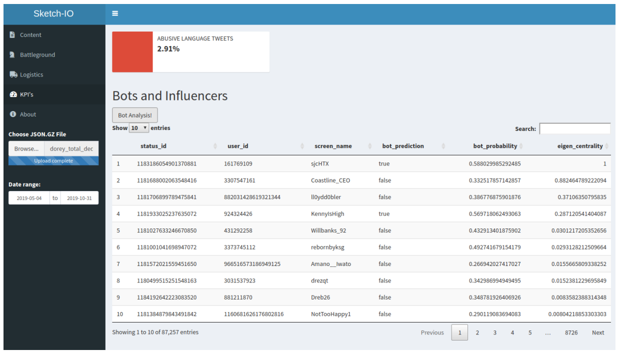
The next two analytics provided on this tab are focused on helping the analyst understand the communities and topic groups found in the data. Communities are groups of connected individuals that often gather virtually around a common interest or group of related topics. Topic groups are groups of people who discuss a common topic, but are not necessarily connected in the virtual or physical environments. To identify communities, we use the Louvain community detection algorithm [46] to detect communities in the conversation network, and then identify common hashtags in the 10 largest communities. To conduct topic group analysis, we concatenate hashtags by user and use this as a corpus for Latent Dirichlet Allocation (LDA) analysis for identifying topic groups [45]. By default this breaks it down into 5 topic groups. We acknowledge that aggregating all social media by user can create significant noise, especially when a user is involved with a wide variety of topics. This compromise facilitates reasonable computation times. Using concatenated hashtags as tokens, however, can only be meaningful if we have more than one or two tokens per document. This was accomplished by concatenating hashtags by user.

7.5.4 Finding Logistics Functions

As discussed above, IO Campaigns leverage the virtual world to access funding and to organize and coordinate meetings and protests. To operationalize this, Sketch-IO searches for URLs and other key words associated with these online funding mechanisms (GoFundMe, Paypal, Bitcoin, etc) and organizational tools (meetup.com, etc). These are provided to the analyst in a searchable table.

7.5.5 Operationalizing Key Performance Metrics

The tab labeled *KPI's* allows an analyst to explore key performance indicators for the campaign. A screen capture of this is provided in Figure 7.10.



Bots and Influencers

	status_id	user_id	screen_name	bot_prediction	bot_probability	eigen_centrality
1	1183186054901370881	161769109	sjcHTX	true	0.588029985292485	1
2	1181688002063548416	3307547161	Coastline_CEO	false	0.332517857142857	0.882464789222094
3	1181706899789475841	882031428619321344	ll0ydd0bler	false	0.386776875901876	0.37106350795835
4	1181933025237635072	924324426	KennyIsHigh	true	0.569718062493063	0.287120541404087
5	1181027633246670850	431292258	Willbanks_92	false	0.432913401875902	0.0301217205352656
6	1181001041698947072	3373745112	rebornbyksg	false	0.492741679154179	0.0293128212509664
7	1181572021559451650	966516573186949125	Amano__Iwato	false	0.266942027417027	0.0155665809338252
8	1180499515251548163	3031537923	drezqt	false	0.342986994949495	0.0152381229695849
9	1184192642223083520	881211870	Dreb26	false	0.348781926406926	0.0083582388314348
10	1181384879843491842	1160681626176802816	NotTooHappy1	false	0.290119083694083	0.00804218853303303

Showing 1 to 10 of 87,257 entries

Figure 7.10: The KPI tab looks at measures of how the conversation is going. This includes measuring abusive language, the influence of bots in the conversation, and the celebrities that are propagating bot messaging.

The prototype dashboard provides three analytics of interest. The value box at the top of the page shows the analyst the percentage of tweets that contain abusive or profane language. This measure indicates how effective the IO campaign has been in creating a divisive conversation.

The next analytic is an interactive table that simultaneously provides a Tier 1 bot score and eigenvector centrality measure of influence (based on the communications network). By exploring and sorting this table, and analyst can identify bots and other IO campaign accounts that have achieved high influence as measured by network centrality. Eigenvector centrality was selected because it helps identify those accounts that are connected to important accounts.

The final table identifies celebrity accounts that retweet bots. “Celebrity” status is determined by any account that has over 50K followers. To identify bots, this algorithm extracts the retweeted status and user object from retweet tweets, writes these original tweets and user objects to disk, and then runs bot-hunter Tier 1 on the resulting data to determine how likely an account is a bot. Retweets of verified accounts are removed. Only celebrity retweets of likely bots that are not verified are returned to the user for interactive exploration.

7.6 Future Work

There are various other approaches one could take to map information warfare campaigns. One approach would involve mapping the information environment to lines of effort, as indicated in Figure 7.11. While our topic group analysis takes a step in this direction, it still falls short of fully delineating campaign lines of effort in multi-media. Another method would be to fill in the template of WHO says WHAT to WHOM and with what EFFECT. This tool would largely focus on attribution of perpetrator and attribution of the target audience, with some content summarization and campaign assessment. While somewhat overlapping with Sketch-IO, both of these approaches would provide a different perspective on a given IO campaign. Due to scoping limitation, these were not explored in depth by our current effort, given our limited focus on attribution and content summarization.

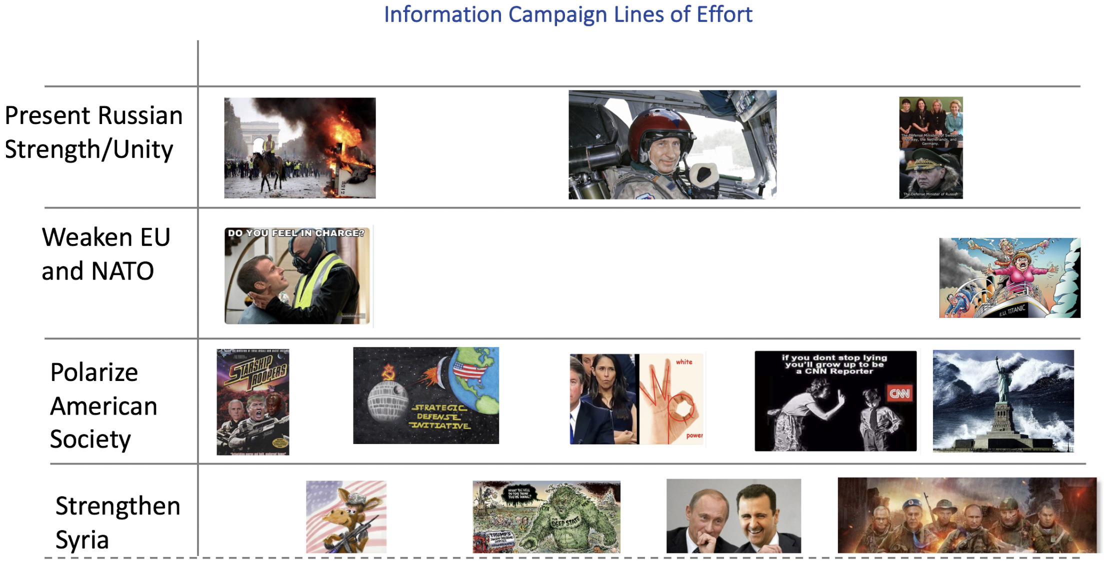

Information Campaign Lines of Effort
Present Russian Strength/Unity
Weaken EU and NATO
DO YOU FEEL IN CHARGE?
Polarize American Society
STARSHIP TROOPERS
STRATEGIC DEFENSE INITIATIVE
white
power
If you dont stop lying you'll grow up to be a CNN Reporter
CNN
Strengthen Syria
DEEP STATE

Figure 7.11: Russian Disinformation Lines of Effort.

7.7 Conclusion

Social cyber security practitioners require a framework for describing general information campaigns that can be operationalized to quickly analyze and describe a specific information campaign. Sketch-IO provides this general framework to help senior leaders understand the mutually supporting parts of a campaign as well as providing analysts a means to disassemble and understand the pieces and parts of a given information campaign.

Appendix A

Characterization and Comparison of Russian and Chinese Disinformation Campaigns

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

A.1 Introduction

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

One of the actions that Twitter took is to suspend accounts associated with state sponsored-propaganda campaigns and then release this data to the public for analysis and transparency. So far they have only released data associated with state-sponsored manipulation and no other actor types. A summary of the data that they released is provided in Table A.1. The largest and most prominent of these is the data associated with the Russian Internet Research Agency (IRA) and the Chinese data. The IRA data includes a well-documented information campaign to influence an election and otherwise cause division in the United States, and the Chinese data is associated with information manipulation around the Hong Kong protests.

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

Table A.1: List of Datasets that Twitter has released in association of state sponsored information manipulation.

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

The IRA data that we will study in this paper is the original data set that Twitter released under their then nascent elections transparency effort. This release was spurred by the fall-out after the 2016 US election and increasing evidence of Russian manipulation. The data has been studied as part of the Mueller Special Counsel investigation as well as several independent analyses conducted on behalf of the US Senate.

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

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

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

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

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

In order to characterize and then compare these data sets, we will develop and illustrate the use of social cybersecurity analytics and visualization. In this paper we will specifically focus on visual network analysis, new geographic analysis using flag emojis, temporal analysis of language and hashtag market share, bot analysis using several supervised machine learning models, meme analysis of image memes, and analysis of state-sponsored media involvement. We will then finish up by analyzing and discussing the number of accounts in the periphery data that are still conducting or supporting state-sponsored information manipulation. Research such as this is key for threat assessment in the field of social cybersecurity [179].

A.2 Literature Review

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

A.2.1 Russia Internet Research Agency Data

Russia's Internet Research Agency (IRA) is a St. Petersburg based company that conducts information operations on social media on behalf of the Russian government and businesses. The company began operations in 2013 and has trained and employed over 1000 people [83].

The IRA data has had more time and research effort than the newer Chinese manipulation data. Even before Twitter released the data to the public they allowed several research organizations an early analysis to accompany the release. Notable among these preliminary and largely exploratory analysis is the research by the Digital Forensic Labs [183].

The Special Investigation "Mueller" report, released on April 18, 2019, detailed the IRA operations [176]. The 443 page report contains 16 pages dedicated to IRA manipulation of information surrounding the 2016 US Presidential election. The manipulation detailed in the redacted report includes organization of grassroots political efforts and use of accounts masquerading as grass roots political efforts. The report indicates that the IRA accounts posed as anti-immigration groups, Tea Party activists, Black lives matter activists, LGBTQ groups, religious groups (evangelical or Muslim groups), as well as other political activists. It also detailed the methods used and the organization of personnel against these methods. Two IRA employees received visas and traveled to the United States in order to better understand the social, cultural, and political cultures. IRA employees operated accounts initially focused on Twitter, Facebook, and Youtube but eventually including Tumblr and Instagram accounts. It also details the purchase of advertisements. It details a separate bot network that amplified IRA inauthentic user content. It noted that celebrities, politicians, and news outlets quoted, retweeted, or otherwise spread IRA messaging. The report outlines throughout the 16 pages how messaging for Trump was positive and supportive while the messaging for Clinton was negative. The IRA was also central to the February 2018 indictment of 13 Russian nationals by Special Counsel Robert Mueller [15].

The second report regarding the IRA was conducted by New Knowledge at the request of the US Senate Select Committee on intelligence (SSCI) and focused on Facebook, Instagram, Twitter, Youtube, Google+, Gmail, and Google Voice involving the IRA. The report also shows some evidence of IRA activity on Vine, Gab, Meetup, VKontakte, and LiveJournal. The data that Twitter provided to New Knowledge was roughly the same data that was released to the public, but was not hashed and contained IP addresses and other information. This highlights the IRA switch from Facebook/Twitter to Instagram following their negative publicity. It highlighted that Instagram outperformed Facebook, highlighting the importance of images and memes in information operations. Like the Mueller report it highlights targeted communities. It also discusses voter suppression operations, such as encouraging voters to vote for a third candidate, stay home on election day, or false advertisements for voting on Twitter. In addition to highlighting pro-Trump and anti-Clinton campaigns, it also highlights activity meant to divide, such as secessionist messaging. It then conducts temporal analysis, URL analysis, and other content analysis. They highlight some of the tactics, branding, and recruitment. It also highlights the IRA's attacks against Republican primary candidates. They conduct extensive analysis of the memetic warfare. They highlight the IRA tactic of amplifying conspiracy theories. Finally, they thoroughly highlight efforts to divide America through secession ("if Brexit, why not Texit"). To summarize, their analysis was primarily content, strategy, and effects across a sophisticated campaign that targeted Black, Left, and Right leaning groups. [83]

The Computational Propaganda Project, like New Knowledge, was provided data by the US Senate Select Committee on Intelligence, to include the Twitter IRA data. In addition to temporal analysis, categorical analysis, target population identification, limited network analysis, hashtag and content analysis. It focused on cross-platform activity [131].

Several other notable research efforts on the IRA include Arian Chen's lengthy New York Times Article entitled "The Agency" which details how the IRA organizes false alarms such as their Columbian Chemicals Explosion Hoax and the Ebola virus hoax [63]. Badawy et al conducts research of the 2016 IRA data and analyzes to what extent the effort supported the political left versus the political right [20], and is probably the closest article to the effort that we propose. Note that the Badawy effort only focuses on IRA data, and does not include any discussion of the Chinese data.

A.2.2 Chinese Manipulation of Hong Kong Narrative

In August 2019 Twitter released data associated with information and platform manipulation by the Chinese government around the Hong Kong protests. Twitter claims this was a state-backed information operation. As evidence for this claim, they point to the fact that all of the activity and the associated IP addresses on the suspended accounts originated from within the People's Republic of China (PRC) even though Twitter is blocked by the PRC (i.e. China's 'Great Firewall'). While some users in China access Twitter through VPNs, the nature of VPNs means the IP addresses aren't from within the PRC. Twitter suspended the accounts for violating terms of service [203]. Censorship, while well documented, is difficult to measure [90].

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

While there is limited reporting on this data, we do want to call attention to the most thorough analysis we've found to date. The most comprehensive analysis we've found was conducted by Uren et al at the Australian Strategic Policy Institute [238]. This research highlights that these accounts attacked political opponents of the Communist Party of China (CPC) even before they began influencing the events in Hong Kong. Some of the primary conclusions of the report is that the Chinese approach appears reactionary and somewhat haphazard. They did not embed in virtual groups and slowly build influence, but rather generated simple spam that supported their messaging. This report does go into extensive temporal and geographic analysis that we will at times enhance but not duplicate. They do highlight that the lack of sophistication may be because it was outsourced to a contractor or because the government agency overseeing the operation lacked a full understanding of information operations. This report also highlights the fact that many of these accounts appear to be purchased at some point in their history. The authors show that 630 tweets contain phrases like 'test new owner', 'test', 'new own', etc. which are commonly used to show that a given account has come under new ownership.

A.3 Data

Twitter is a core platform for the global conversation, providing an open market for opinions and beliefs. By 2014 Twitter surpassed Facebook citations in the New York Times and by 2016 the New York Times cited Twitter more than twice as much as Facebook [243]. Online media often include Twitter posts of celebrities, politicians, and other elites in their content. To some extent, Twitter captures more of the global conversation (particularly in the West) while Facebook captures more of the local and topical conversations. Given this important opinion market, numerous actors attempt to market their ideas and at times manipulate the marketplace for their benefit.

Table A.2: Summary of Data

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

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

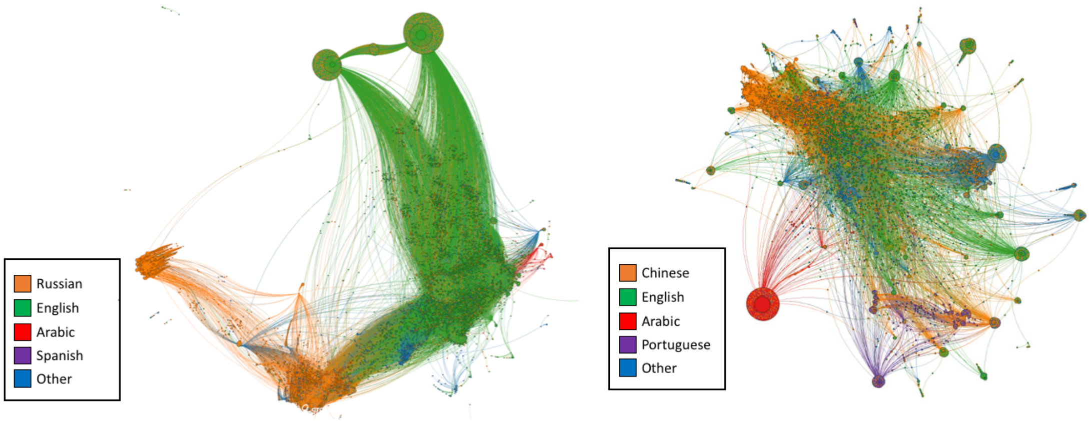
A.4 Characterization and Comparison

A.4.1 Network

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

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

The conversation in the accounts used by the Chinese information operations is more complex primarily since these accounts seem to be recently purchased by the Chinese government or government proxy, and the earlier histories of these accounts is varied. We observe that, even though Arabic and Portuguese have a large proportion of the conversation by volume, their use is relegated to a few accounts that are structurally segregated from the rest of the network. The



(a) Russia Core Conversation

(b) China Core Conversation

Figure A.1: The conversational network of the core accounts suspended and released by Twitter (colored by most recent language used by account).

Chinese and English language campaigns are much more intertwined as China directs their information campaign at the Western world and at Hong Kong. While the messaging is aimed at Hong Kong, it is not necessarily aimed internal to China since Twitter is blocked by China's firewall.

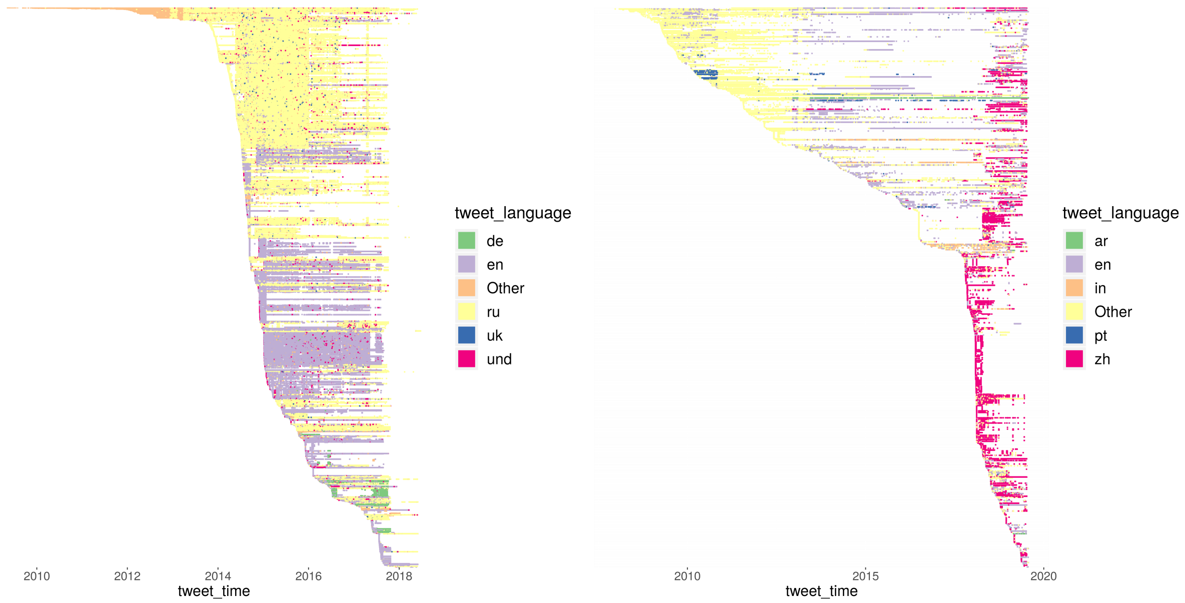
A.4.2 History of Accounts

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

In the case of the Russian IRA, the timeline demonstrates a persistent effort to embed in both Russian and English language societies. Specific accounts embedded into target cultures and subcultures, learned to interact within the values and beliefs of the subculture, and then began to manipulate both the narrative and the network in these subcultures. We do see some evidence of dormancy with some accounts leaving the conversation for sometimes years at a time, but nonetheless consistent effort to curate virtual personae within a narrow context.

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

The history of the accounts shows a very different approach between the two disinformation campaigns. The Russian effort demonstrates a planned and persistent effort to embed into the target society, and especially within target subcultures. They did this in the Russian language to



(a) Russia

(b) China

Figure A.2: Tweets over Time Colored by Language

manipulate their own population, and in English to manipulate beliefs and actions in America. Once embedded these agents continued to develop a following and influence a larger and larger swath of the American populace.

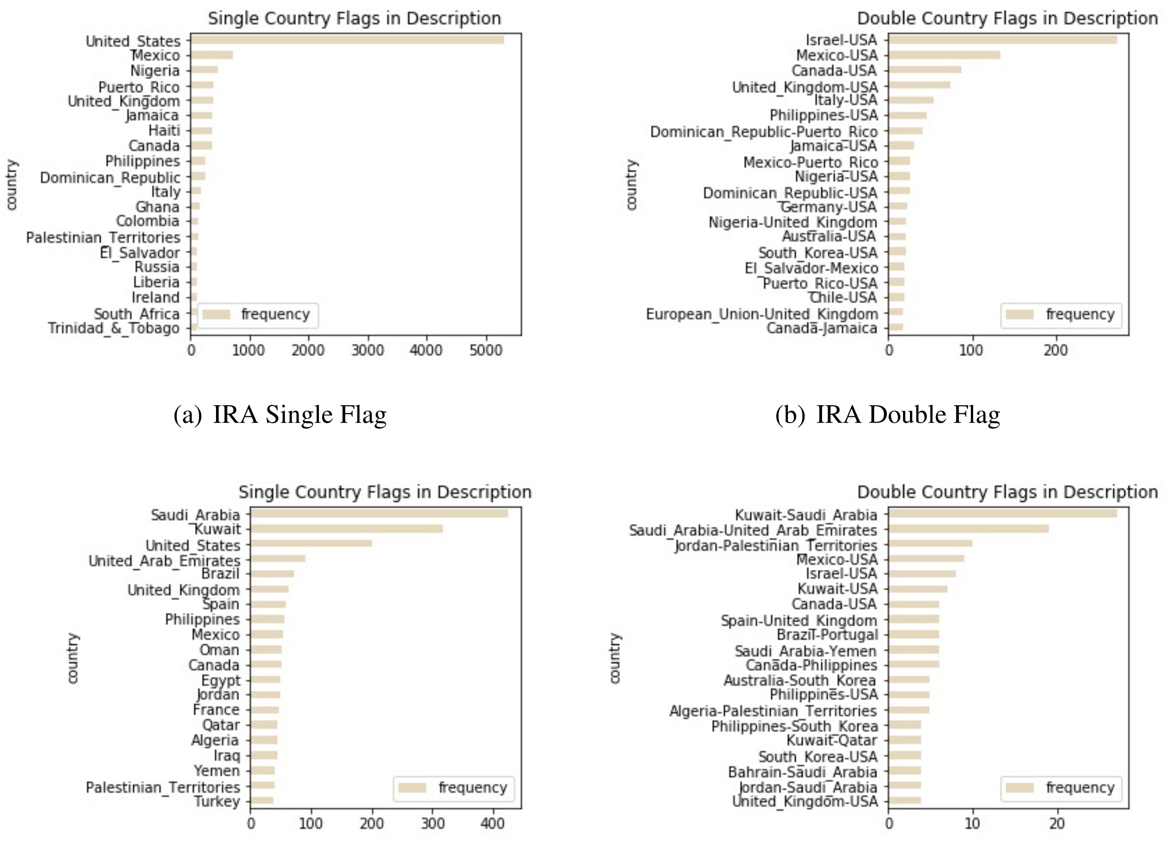
The Chinese approach was much more reactionary, seems less planned, and did not have any persistent effort to embed in networks to affect influence. To some extent, the Chinese effort was simply to spam their narrative across the international community.

A.4.3 Geography of Accounts

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

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

We found evidence of this behavior in the core data set, particularly in the IRA data. Two ex-



(a) IRA Single Flag

(b) IRA Double Flag

(c) China Single Flag

(d) China Double Flag

Figure A.3: The Distribution of Flag Emoji's in Account Descriptions. The high volume of unexpected flags used for the China data (such as Kuwait/Saudi Arabia) is due to the fact that many of these accounts were recently purchased by the Chinese government, and therefore most tweets and account descriptions by these accounts were produced by their previous owners.

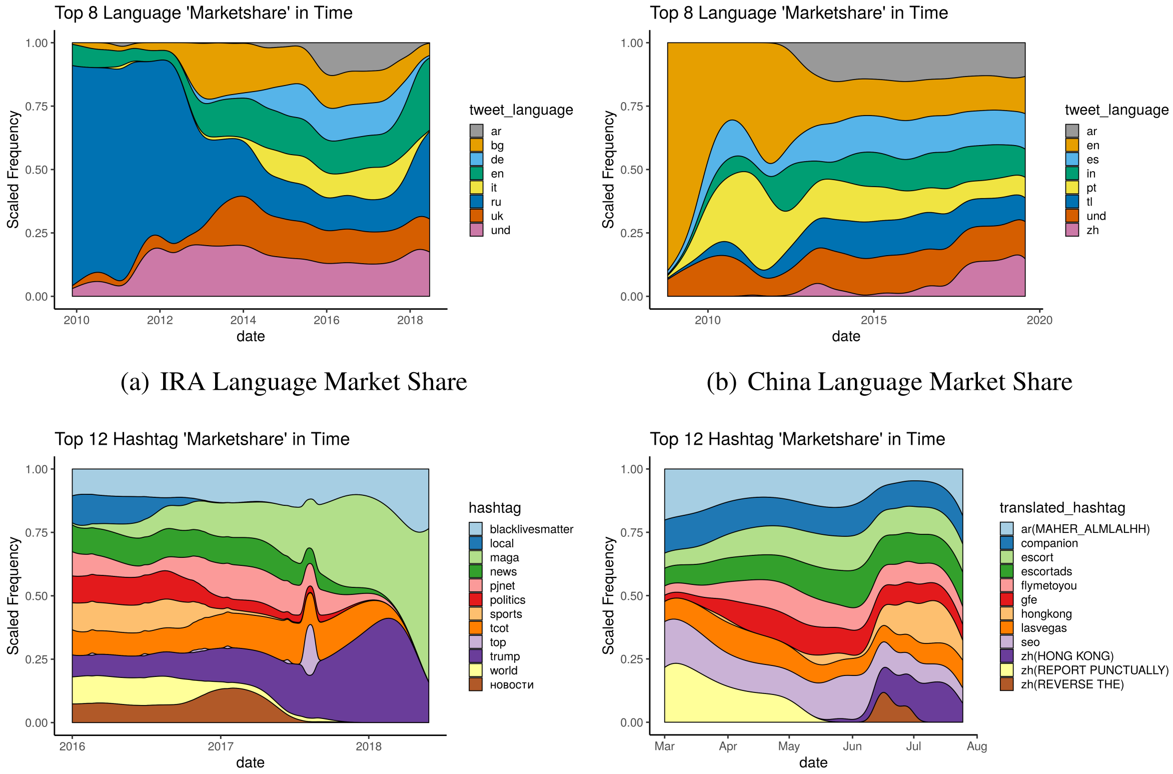
amples are 🇷🇺 in 🇺🇸 and Russian 🇷🇺 living in the US🇺🇸 . In these cases, a description like this allows the casual observer to rationalize why the account switches back and forth between Russian and English and between Russian social/political conversations and American social/political issues.

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

A.4.4 Calculating Content Marketshare Over Time

Although we've already looked at the histories of these accounts, we wanted to understand temporal distributions better so that we can understand how these accounts were used over their life span as well as in the world events they're respectively associated with. To do this we explored the use of language and content over time with temporal market share.

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



(a) IRA Language Market Share

(b) China Language Market Share

(c) IRA Hashtag Market Share

(d) China Hashtag Market Share

Figure A.4: Normalized Marketshare of Language and Hashtags for Core IRA and Chinese data suspended by Twitter

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

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

A.4.5 Bot Analysis

Social media bots are any account that has some level of action being automated by a computer. On Twitter tweeting, retweeting, replying, quoting, liking, following, and general searching can all be automated. In this section we leverage several bot detection tools to predict the number of accounts that appear to have automated behavior. Memes and bots are tools used to conduct information maneuvers in influence campaigns [34].

The models used below are two external models as well as two that were developed by our team. The first external model is the Debot model [60]. The Debot model is an unsupervised model that finds bots that are correlated using warped correlation. In other words, this model finds two or more accounts that are posting the same content at roughly the same time. The Debot team continually monitors parts of Twitter, and keeps a database of accounts that they've found to be correlated. In our search through the Russia and Twitter periphery data, we searched the Debot database to identify any of our accounts that have been found before. The second external model is the Botometer model (previously called the BotOrNot model) [74]. The Botometer model is a supervised machine learning model with well-documented feature space. The Botometer Application Programming Interface (API) accepts a user ID or screen names as input, scrapes the Twitter API using the consumer provided keys on the server side, and then returns a score for content, friends, network, sentiment, temporal, user, and the universal score for the account. Given this method, Botometer scores are only available for accounts that are still active (i.e. not suspended, private, or otherwise shutdown). Due to the time required to scrape the timeline, in both of our data sets we randomly sampled 5,000 accounts for the Botometer model.

We've also listed scores for two models developed internally. The Bot-Hunter suite of tools provides supervised bot detection at several data granularities. Tier 1 conducts bot detection with a feature space developed from the basic tweet JSON data that is returned by the Twitter API [30]. This includes features extracted from the *user* object and the *tweet* object. Tier 2 performs bot detection using the user's timeline (adding more content and temporal features), and Tier 3 uses the entire conversation around an account to predict the bot score [33]. Due to the computational cost of running Tier 3 (approximately 5 minutes per account), it is best for only a handful of accounts and was not used on these data sets. The Bot-Hunter Tier 1 model was run on all data, and the Tier2 was run on a random sample of 5000 accounts. Note that unlike Botometer, Bot-Hunter runs on existing data and was therefore able to predict on core, periphery, and suspended accounts. We've also developed an abridged version of Bot-Hunter Tier 1 that can run on the core data since it doesn't contain all features available for the unabridged model.

From Table A.3 we see that models predict that 9-15% of the Russian core and periphery have likely automated behavior, with Hong Kong estimates slightly lower with Bot-Hunter predicting

Table A.3: Bot prediction for *core* and *periphery* data (% of Total).

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

8-14% automated behavior and Botometer as the outlier with 28% prediction.

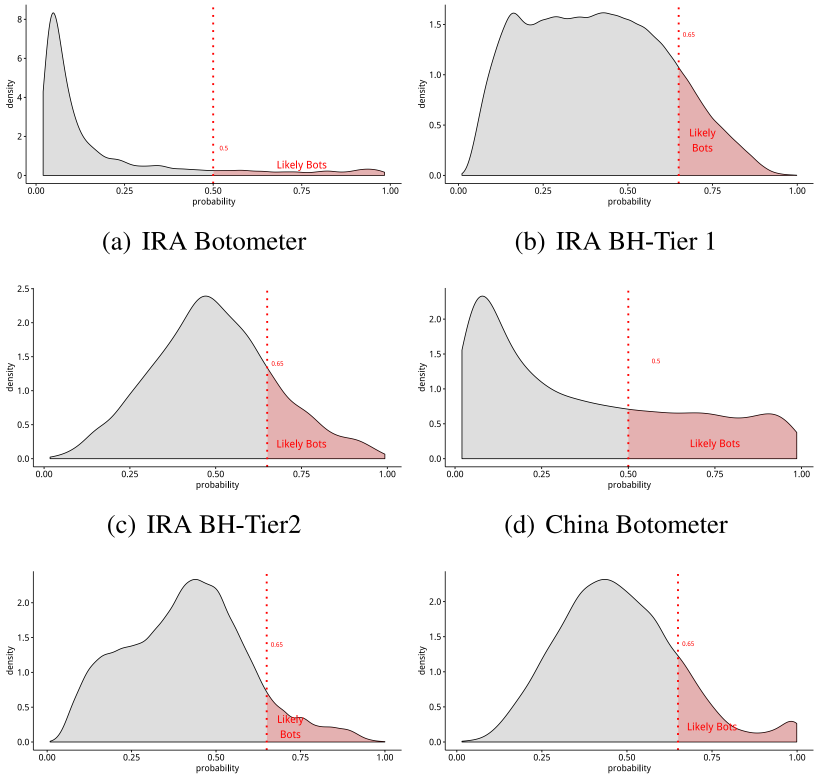
We get even more insight into these models and data by looking at Figure A.5. This shows the probability distribution and chosen thresholds for each of the models on the periphery data. The biggest takeaway in these images is the difference between the shape of the Botometer model and the Bot-Hunter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

A.4.6 Multi-Media Analysis

Richard Dawkins originally created the word meme in his book Selfish Gene in which he defined a meme as a "...noun that conveys the idea of a unit of cultural transmission, or a unit of imitation" [77]. Shifman later adapted and defined internet memes as artifacts that "(a) share common characteristics of content, form, and/or stance; (b) are created with awareness of each other; and (c) are circulated, imitated, and transformed via the internet by multiple users" [215, 216].

Internet memes, particularly multi-media memes, are increasingly used in online information warfare. This phenomenon has been highlighted in articles like the New York Time "The Mainstreaming of Political Memes Online" [49], and has been dubbed memetic warfare. To analyze memes in these two data sets, we developed a deep learning meme classifier to extract memes from the multi-media archives that Twitter shared along with the data. We ran this classifier on all images in the IRA data set, and on all Hong Kong related images in the China data set. Examples of IRA memes are provided in Figure A.6 and examples of China memes are provided in Figure A.7.

From our analysis of these we see the IRA use memes at a much higher volume and sophistication. IRA memes involve significant creative content development and solid understanding of target subculture and biases. The IRA memes uses standard meme templates so that their memes fit in with the deluge of internet memes flowing around an election event. Their memes also cover the full spectrum of information operations forms of maneuver, to include all aspects of



(a) IRA Botometer

(b) IRA BH-Tier 1

(c) IRA BH-Tier2

(d) China Botometer

(e) China BH-Tier 1

(f) China BH-Tier 2

Figure A.5: Probability distributions for bot prediction for Botometer, Bot-Hunter(BH) Tier 1 and Tier 2 with threshold shown.

supporting or attacking narratives and supporting or attacking networks.

The Chinese memes, in contrast, were hastily created and are primarily informational in nature. In fact, in many respects they do not meet the definition of a meme that Dawkins and Shifman put forward above, since we do not see significant evolution and transformation by multiple users. To some extent, they represent another facet of a campaign to spam the world with a particular narrative about the Hong Kong protests. Across their information campaigns, the Chinese seem reluctant to uses memes. While part of this may be cultural, another reason for their reluctance may be a worry that the evolution and propagation of memes is in the hands of the masses, not tightly controlled by central authorities. Memes can quickly turn negative toward the Chinese Communist Party and its leadership, as they did with Winnie the Pooh memes, causing party leadership to ban and censor Winnie the Pooh memes [169].

A.4.7 State Sponsored Accounts

Part of the role of bots within information operations is to amplify certain voices within the conversation. With state-sponsored information operations, this often means amplifying state-sponsored media. In recent years, Russia has increased the worldwide penetration of RT, Sputnik, and other state-sponsored news agencies, while China has been gaining greater international

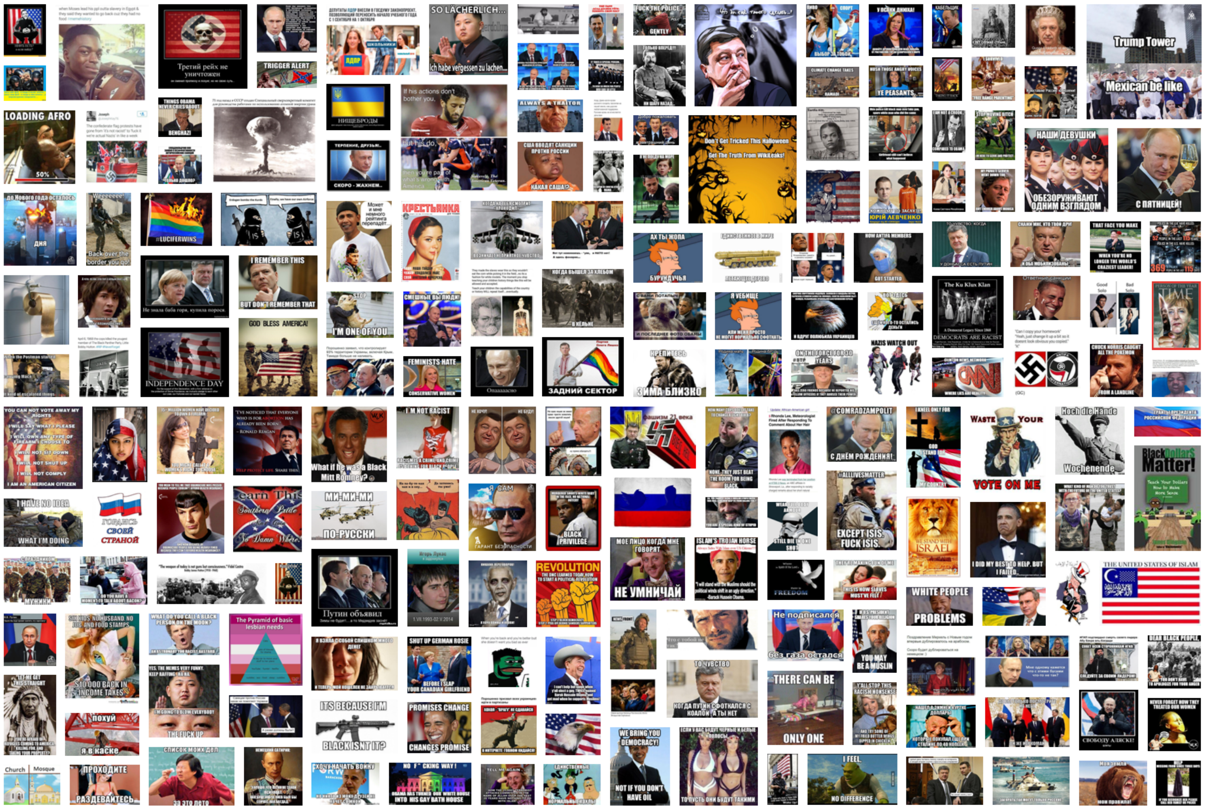


Figure A.6: Russian IRA Memes

penetration with China Xinhua News. To measure the extent that to which this data is amplifying these voices, we collected a large list of all Twitter handles associated with these Russian and Chinese state owned media companies, as well as handles associated with several other countries state owned media (for example the US Voice of America) for comparison. While the degree to which each of these handles spread state "propaganda" varies widely, we provide them for comparison.

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

Table A.4: Add caption

State Owned Media	IRA Core Data	Chinese Core Data
Russian	72,846	8
Chinese	226	1,400
American	11	0
Korean	0	2
German	62	2

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

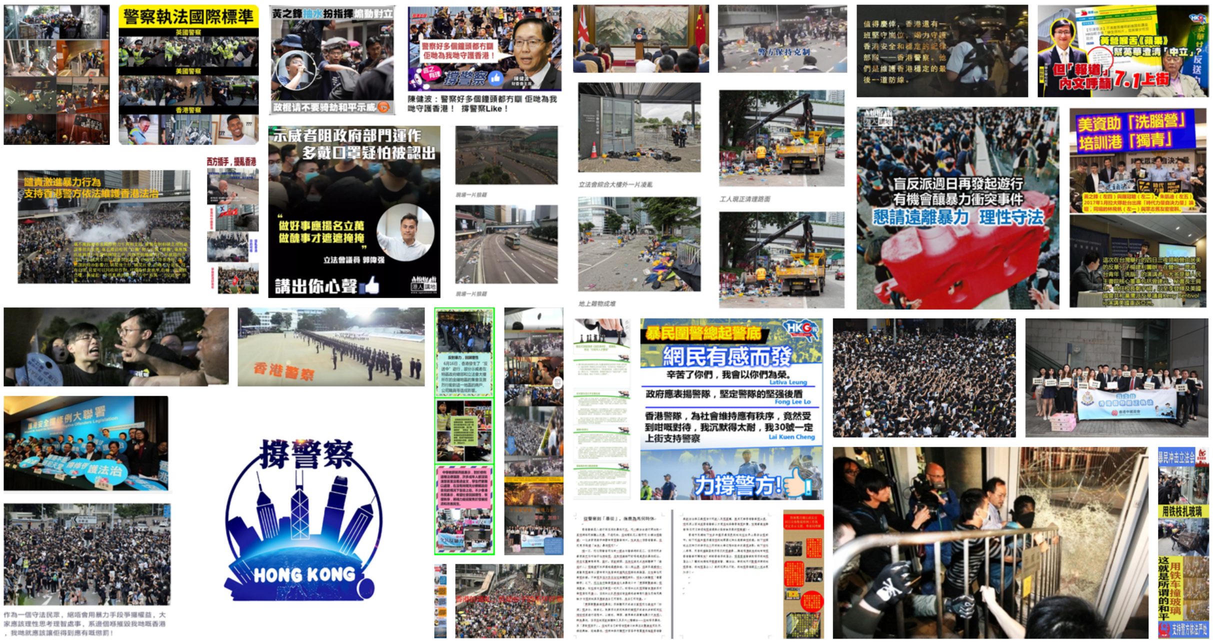

警察執法國際標準
英國警察
美國警察
香港警察
黃之鋒抽水扮指揮煽動對立
政棍请不要骑劫和平示威
警察好多個鐘頭都冇瞓 佢哋為我哋守護香港！
撐警察
陳健波：警察好多個鐘頭都冇瞓 佢哋為我哋守護香港！ 撐警察Like！
警方保持克制
信得慶幸，香港還有一班堅守崗位，竭力守護香港安全和穩定的紀律部隊——香港警察，他們是維護香港穩定的最後一道防線。
但「報道」內文呼籲7.1上街
示威者阻政府部門運作 多戴口罩疑怕被認出
做好事應揚名立萬 做醜事才遮遮掩掩
立法會議員 郭偉强
講出你心聲
譴責激進暴力行為 支持香港警方依法維護香港法治
西方插手，擾亂香港
立法會綜合大樓外一片凌亂
工人現正清理路面
地上雜物成堆
盲反派週日再發起遊行 有機會釀暴力衝突事件
懇請遠離暴力 理性守法
美資助「洗腦營」培訓港「獨青」
香港警察
撐警察
HONG KONG
暴民圍警總起警惕
網民有感而發
辛苦了你們，我會以你們為榮。
政府應表揚警隊，堅定警隊的堅強後盾
香港警隊，為社會維持應有秩序，竟然受到咁嘅對待，我沉默得太耐，我30號一定上街支持警察
力撐警方！
用铁栈扎玻璃
这就是所谓的和平

Figure A.7: China Memes

A.5 How many similar actors are left?

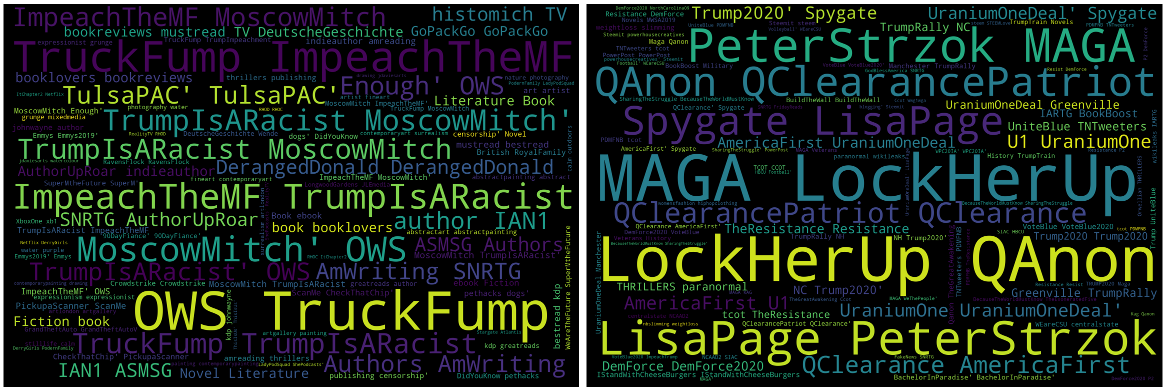
One of the biggest questions that remains after going through this data is "How many state-sponsored actors are still at large in the virtual world and currently manipulating world events?" To try to answer this question we spent some time analyzing the *periphery* data that is still mostly 'alive' and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

As shown in Table A.3, approximately 10% of both streams exhibit bot like behavior (these are again conservative estimates). Of the accounts in the periphery, 85.2% of the Russian accounts and 64% of the Chinese accounts are active, meaning they are not dormant and have tweeted in the last 6 months. Additionally, these accounts continue to amplify state-owned propaganda. The IRA *periphery* amplifies Russian state-owned media 6,023 times, and the China *periphery* amplifies Chinese state-owned media 1,641 times.

Below we try to capture the primary topics that these accounts are embedding in. To do this we sampled 5000 accounts from the periphery of Russia and China, collected the last 200 tweets associated with each account. After selecting only those tweets in the last 6 months, we conducted topic analysis with latent Dirichlet allocation (LDA). By optimizing the Calinski

Harabaz score, we chose a k of 10 for LDA.

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



(a) Possible Russia influence in the US Left (8-10% of Accounts)

(b) Possible Russia influence in the US Right (10-12% of Accounts)

Figure A.8: Current Information Operations by Russia found in the *periphery* data

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

A.6 Conclusion

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

- The IRA's effort included identification and study of target subcultures with significant effort to shape messaging to leverage existing biases.
- The Chinese effort was aimed at Hong Kong and the international community at large without evidence of extensive effort to identify a target audience or craft messaging for a specific audience.

- The IRA effort demonstrates an understanding of internet memes and a willingness to take risks in releasing multi-media messaging that will evolve in the masses.
- The Chinese effort demonstrates an unwillingness to release internet memes that will evolve outside of the direct control of central authorities.
- Both efforts, but particularly the Russian effort, demonstrate an effort to use these covert information operations to enhance the overt information operations conducted by state-owned media companies.

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

This work lays the foundation for building a repeatable end-to-end process for characterizing malicious actors in disinformation streams in social media, which is essential for national security [35]. These efforts to characterize actors will assist social cybersecurity analysts and researchers in getting beyond the binary classification of ‘bot or not.’ Future research will describe and illustrate this full workflow and several different data sets.

Appendix B

Agent Based Simulation of Bot Disinformation Maneuvers in Twitter

Multiple state and non-state actors have recently used social media to conduct targeted disinformation operations for political effect. Even in the wake of these attacks, researchers struggle to fully understand these operations and more importantly measure their effect. The existing research is complicated by the fact that modeling and measuring a person's beliefs is difficult, and manipulating these beliefs in experimental settings is not morally permissible. Given these constraints, our team designed an Agent Based Model called `twitter_sim` that allows researchers to explore various disinformation forms of maneuver in a virtual environment. This model mirrors the Twitter Social Media Environment and is grounded in social influence theory. Having built this model, we demonstrate its use in exploring two disinformation forms of maneuver: 1) *backing* key influencers and 2) *bridging* two communities.

B.1 Introduction

As more information has become available detailing how Russia's Internet Research Agency (IRA) manipulated various societies and political processes across the world, researchers have worked to document the IRA's methods [34], develop ways to detect these methods [30, 60, 95], and determine how effective these methods are [39].

While these research efforts have effectively identified methods, target audiences and developed models to detect certain agents (bots/trolls), they still struggle to answer the question of impact. The primary challenge being that it is difficult to measure and label a person's beliefs, and even harder to measure the evolution of these beliefs over time while enumerating resulting decisions/actions. If is often difficult for a person to identify how a myriad of real and virtual social interactions and messaging affect their own decisions, let alone someone else to measure these. Laboratory experiments that attempt to manipulate a person's beliefs are not morally permissible, therefore, human subject experimental studies provide limited utility.

For decades researchers have attempted to fill this gap with models of information diffusion, rumor propagation, and social influence more generally. These models include systems dynamics, discrete event simulation, and more recently agent based models. While others have modeled Twitter in these simulations, authors have generally abstracted away the mechanics of the Twitter environment (Tweets, replies, retweets, mentions, following, news feeds, etc). In our effort to model and evaluate disinformation forms of maneuver, we felt that it was important to explicitly model the mechanics of the Twitter environment, and we have not found any research effort that has done this. Our team set out to develop a special purpose ABM called *twitter_sim* where we model a given social media environment (Twitter), insert malicious agents (bots/trolls), conduct various disinformation forms of maneuver, and evaluate the emerging behavior. In addition to exploring disinformation maneuvers as we have done, *twitter_sim* could be used for modeling marketing campaigns or as the backbone for a virtual social media training environment.

`twitter_sim` includes most of the actions of the Twitter environment (tweet, reply, retweet, mention, follow). It also includes heterogeneous human behavior on the platform, including varied rates of access, limited attention, dynamic network links based on homophily and social cues, and changed beliefs based on level of exposure weighted by homophily and authority.

In addition to describing and validating the *twitter_sim* model, we have applied it on simulated networks to explore two disinformation forms of maneuver used by bots/trolls: 1) *backing* key influencers and 2) *bridging* two communities. Backing involves following and retweeting influencers to amplify their message. Bridging involves building links between two communities in order to introduce the ideas of one into the other. In this exploration our central question is 'What emergent behavior do we observe when bots back influencers or bridge networks in social media?' These maneuvers were selected because they were used by the IRA in information operations in the United States and Ukraine, and because these two methods can be similar in practice (bridging often involves backing influencers simultaneously in two communities in an effort to bring them closer).

B.2 Related Works

Early models of information diffusion on social media tried to use an epidemiology model known as the SIR Model (Susceptible, Infected, Recovered). Daley and Kendall produced the first model based on this [73], which was expanded in the mid 1970's by Maki and Thompson [164]. These models (often referred to as the DK and MK models) describe the three states as ignorants (S), spreaders (I), and stiflers (R). Early models that use the SIR model in social media include [263]. There have since been many different evolutions from the original models, the most prominent including the SI model [79], SIS (susceptible, infected, susceptible) model [125], SIHR (Susceptible, Infected, Hibernators, Recovered) model [267] and the SEIR (Susceptible, Exposed, Infected, Recovered) model [255] . This paradigm is still used in ABM models about Twitter, such as the study of earthquake rumors in Japan [229]. Several other works have criticized the epidemiological approach as overly simplistic since it generally assumes a homogeneous population connected in a simple network with a constant probability of infection, and because recovery mechanisms of epidemics (often vaccination) are different than the infection mechanisms, whereas in rumors the recovery mechanisms are very similar to the infection mechanisms (i.e. anti-rumor messaging) [234]. Li et al provides a thorough general survey of disinformation models [155].

B.2.1 Agent Based Models of Social Media

Several other authors have conducted somewhat similar research using ABM's to model social media, particularly looking at the spread of rumors or misinformation. Tripathy et al [234] provides a network based ABM that consists of Neutral nodes, infected nodes (believed the rumor), vaccinated nodes (believe the anti-rumor before infection, and cured nodes (believe anti-rumor after infection). Additionally, Tripathy et al explores the idea of a *beacon*, which is similar in purpose to the "stifler" in earlier models. They indicate that a given authority detects a rumor at some point after the beginning of the rumor (a time delta that they vary and find important) and then position *beacons* which help to broadcast the anti-rumor message. Serrano [207] adapts earlier models by claiming the *recovered* users will not try to influence their neighbors, offering empirical evidence from Twitter to support this. A recent model by Wang et al. attempts to capture memory, conformity, differences in propensity to produce/spread rumors, and variance in trust to model rumor propagation with information entropy [245].

For those looking for a fuller background into ABM's modeling information diffusion in Twitter, Serrano et al provides a longer survey and analysis [206].

B.2.2 The Construct Model

Some of what we will present in this paper builds on the Construct model originally presented in 1990 by Carley [54] and revisited in 2009 [55]. Construct is a general purpose social influence model that seeks to bring in the complex social dynamics that the are not present in the epidemiological models mentioned above. "Construct is the embodiment of constructuralism, a mega-theory which states that the sociocultural environment is continually being constructed and reconstructed through individual cycles of action, adaptation, and motivation." [55] Construct presents the idea that bounded rationality impact social interaction. Bounded rationality means that agents do not have access to all information (due to social position) and do not process/retain all information that they do have access to. The beliefs of others impact an agent's beliefs through a process called social influence [102].

Within Construct, the likelihood of interaction is based on relative similarity and relative expertise. Construct agents have general and transactive memory. General memory is the facts it knows and the beliefs it holds. The transactive memory is its view (not necessarily accurate) of "who knows who and who knows what" [54].

While not implementing the full construct model, our model is informed by several concepts that Construct introduces. These include bounded rationality (agents are bounded by position and expertise), likelihood of interaction based on similarity, interactions weighted by similarity and expertise, and a Twitter specific model of general and transactive memory.

A docking lite comparison of *twitter_sim* (our model), Construct, and SIR based epidemiological models is provided in Figure B.1.

Table B.1: Docking Lite Comparison of twitterSim, Construct, and SIR

	twitterSim	Construct	SIR
General Population	✓	✓	✓
Media Agents	✓	✓	
Opinion Leaders	✓	✓	
Information Access	✓	✓	
General Memory	✓	✓	✓
Transactive Memory		✓	
Homophily	✓	✓	
Influence of General Populace	✓	✓	
Limited Attention	✓		
Allow Dynamic Network	✓	✓	

B.2.3 The Twitter Environment

Twitter began in 2006 as a way for people to share Short Message Service (SMS) messages with a maximum length of 140 characters. As such, it quickly became the first and arguably the largest of the "micro-blog" platforms. In the Twitter environment, users have a two way following mechanism that is rather unique among social media platforms. While many platforms allow an undirected network link that is established when both parties decide to be "friends", Twitter allows one-way directed links. Users *follow* other users. This creates a rich network where links represent either one-way following, both users follow each other (and are therefore "friends"), or neither follow.

Within the Twitter environment, users can interact in a variety of ways. These include the following actions:

1. **Tweet:** Users generate a short message that can include multimedia
2. **Retweet:** Users send another user's message to their followers without comment.
3. **Quote:** Users send another user's tweet to their followers with comment (starting a new thread)
4. **Reply:** A user replies to a tweet that someone else makes (remains in same thread as original tweet)
5. **Mention:** A user places another user's screen name to a tweet; the mentioned user is notified
6. **Like:** Users can like a tweet, which increases its prominence on the platform

Another aspect of Twitter and other social networks is that use of Twitter can vary significantly. Some users rarely log on and could be considered dormant. Others use it extremely often, while bots use it at the speed of algorithms. Many models of disinformation on Twitter don't capture this aspect of social media usage, assuming that every user will have a chance of influence in every time step, but this is something that we'd like to explicitly capture.

A Twitter user's feed is only populated by tweets, retweets, and replies produced by those accounts that they follow. While a user can search through the Twitter stream on their own, their feed is only populated by a proprietary algorithm with tweets from those they follow (not those that follow them). *twitter_sim* explicitly models this structural constraint of the environment.

B.2.4 Disinformation Maneuvers

The authors have previously outlined the BEND framework for identifying disinformation maneuvers [34]. In the BEND framework, information operations can target both the information and the network. Often, information warfare architects will attempt to manipulate both at the same time. The BEND Framework is summarized in Table B.2.

Table B.2: Summary of BEND Framework

	Network	Information
Pro	Back Build Bridge Boost	Engage Explain Excite Enhance
Con	Neutralize Nuke Narrow Neglect	Dismiss Distort Dismay Distract

We developed the *twitter_sim* environment primarily to explore these maneuvers in a virtual environment. From the list of 16 maneuvers, we selected *back* and *bridge* to initially explore because of our notion that *bridging* may simply be the same as *backing* but focused on two communities instead of a single community.

B.3 Modeling Twitter Environment with an Agent Based Model

In order to accomplish the goals discussed above, we developed an *twitter_sim* as an Agent Based Model in Python 3.6. Throughout the development process, we leveraged the rich network functions available in the *networkx* package [118]. Building this in Python on top of *networkx* allows the model to be adaptable and scalable (easily run in parallel on large compute resources), while shareable through open source software mechanisms. Although the models in this paper use scale free networks, the *titter_sim* model was developed to easily accommodate experiments on real world networks and events (i.e. humanitarian disaster or election event).

In *twitter_sim* we explicitly model users and their behavior on Twitter. We use three types of users that mirror the DK model: Normal users (ignorants), bots/trolls (spreaders), and truth defenders (stiflers). Only truth defenders and bots aggressively pursue information operations. Normal users, even once their beliefs begin to change, generally do not aggressively engage in a given campaign [207], but do propagate information messages through retweets.

Normal users and truth defenders begin at time 0 embedded into a preexisting network. Bots start at time zero on the periphery of this network with only a single link to a randomly selected node. This therefore explicitly models the challenge that bot creators face in embedding and building position in networks and online communities.

As mentioned above, a Twitter users' feed is only populated by content produced by accounts that they follow. It is not populated with the content of people that follow them. *twitter_sim* models this behavior and populates an agent's Twitter feed with tweets from those they follow (their successors). This means that bots, while producing a much higher concentration of disinformation, will only start having an effect once they become embedded in real networks and build a following.

B.3.1 Twitter use as a discrete event simulation

We have not found a model that considers that Twitter can only influence a user if they log on and read their Twitter feed. The inter-arrival time of people returning to Twitter varies widely. Some people use Twitter multiple times daily, while others check it every other month. Many set up an account, connect with a couple friends, and then never return to the platform.

In *twitter_sim* each agent stores the next time step that the agent will log on to Twitter, and won't read their feed, send tweets or adjust their beliefs until that time step. We've modeled the inter-arrival time as an exponential random distribution parameterized by λ, the mean hourly rate. We varied λ with a uniform distribution ranging from 0.001 (once every 2 months) to 0.75 (18 times per day). These numbers are validated with empirical data later in the paper. Therefore inter-arrival time T is defined as

$$T \sim Exponential(\lambda_i) \quad \text{where} \quad \lambda_i \sim uniform(0.001, 0.75)$$

B.3.2 Limited attention of users

Like other models [251], *twitter_sim* will model the limited attention that users have. During a given session where the user logs onto Twitter, they will only 'read' the last 4 to 20 tweets in their feed. Only read tweets are used in updating a user's beliefs. The number of 'read' tweets is a random uniform integer.

This limited attention behavior means that those accounts with a high in-degree will only read a small portion of their total feed, while less popular accounts have the potential to read most of their feed (depending on the rate of activity for them and their followers).

B.3.3 Homophily impacts decisions

McPhersen et al introduced the idea of homophily in social networks with the idea of "birds of a feather flock together" [170]. McPhersen summarized homophily by stating the "similarity breeds connection." *twitter_sim* measures similarity between agents and uses this similarity to create new links as well as weight information (i.e. tweets from agents that are more similar to me will have greater impact on my beliefs).

In our model, homophily, or similarity, is measured by the similarity of out-links (followed accounts). If two agents follow many of the same accounts, then they are more similar. The overlap of followed accounts is measured by a Jaccard similarity of the adjacency matrix and is updated on a weekly basis. The Jaccard similarity of User A and User B is therefore computed as follows:

$$similarity = \frac{successors_A \cap successors_B}{successors_A \cup successors_B}$$

B.3.4 Influence

The level of influence or prestige for a user is measured as the normalized in-degree of the user. Other models have also used influence to inform probability of acceptance [245], though in our case we use only in-degree since Twitter is by nature a directed network. Most people would agree that 'important' or 'popular' accounts have many followers, but don't necessarily follow many accounts themselves.

The number of accounts a user can follow in Twitter is artificially capped at 5,000 until an account has more than 5,000 followers. In our scale-free networks out-degree is limited to a given percentage of the overall nodes in the network (usually 10-20%). In-degree is unconstrained.

B.3.5 Mentions and Retweets

Twitter allows users to *mention* a user in a tweet. This is done for multiple reasons, and alerts the mentioned user of the tweet. Mentions can also be used by average users as well as bots/trolls in an attempt to gain followers. Our model produces mentions with a given probability, and then a small portion of new links are directed to mentions.

Twitter allows a user to *retweet* another user's tweet. Retweeting without adding additional comment primarily serves the purpose of propagating the message. In our model all agents retweet with a given probability. All retweets carry the value of the original tweet (weighted by the homophily and influence of the originator) as opposed to the retweeter.

B.3.6 Stiflers

On Twitter users actively counter disinformation, and this is often done with *quotes* and *replies* [19]. These users are often referred to as *stiflers* or *beacons*. 10% of the users in *twitter_sim* are labeled as *stiflers*. These users actively combat disinformation by spreading the counter message, which has a negative affect on disinformation belief.

Stiflers send one reply for every disinformation retweet they read in their inbox. This means that stiflers will be constrained by bounded rationality, and will only be able to counter messages that they are aware of (meaning they are produced by neighbors of the stifler). They are also constrained by limited attention since they only counter tweets they read.

B.3.7 Modeling Impact of Global Conversation

Users are not only influenced by content produced by the accounts they follow, but are also allowed to search through tweets by topic, user, or hashtag. In doing so they are influenced by the larger *global* conversation occurring on Twitter. Our model explicitly models this as the current mean belief of the network and uses it to update a person's beliefs when they log on. This is captured by the $global_{perc}$ parameter shown below.

B.3.8 Changing Beliefs

Influence is measured as a continuous variable from 0 (does not believe disinformation) to 1 (dedicated disinformation believer). To calculate beliefs, we start by assigning a value to each tweet. This value is calculated as follows:

$$Tweet_{value} = type \times similarity_{ij} \times influence_i$$

where $type \in \{-1, 0, 1\}$ indicating *anti-disinformation* (-1), *noise* (0), or *disinformation* (1). i is the user sending the tweets, and $j \in \{\text{followers of i}\}$. Belief is then computed with

$$belief_t = belief_{t-1} + (mean(tweets_{read}) + global_{perc})) * (1 - belief_{t-1})$$

The final term is designed to constrain the value between 0 and 1. This also causes diminishing returns in belief (i.e. a dedicated disinformation believer will not become an even greater dedicated believer).

B.3.9 Agent Based Model Algorithm

The basic time step of our agent based model is presented in Algorithm 2.

Algorithm 2 Pseudo-code for Twitter Disinformation Agent Based Model

```
initialization
for each time step do
    if start of new week then
        Update similarity matrix
        Update influence vector (in-degree)
        Update global perception
    for each user do
        if If user checks Twitter in this time step then
            Get new wake time
            Read Tweets
            Adjust belief value
            Create tweets
            Add retweets to tweets
            Create mentions
            Send tweets/mentions to followers
            With a given probability create new link with similar user
            With a given probability create new link with mention author
```

The above algorithm is only slightly modified if the user is a bot/troll or stifler. The stifler will send counter-disinformation *replies* instead of tweets. Bots send disinformation mixed with 20% noise, and also attempt to add links during every session (normal users add link with probability 0.05).

B.4 Exploring Known Disinformation Maneuvers

In this section we will use *twitter_sim* to explore emergent behavior when bots/trolls conduct known disinformation forms of maneuver. In our study we decided to model *backing* and *bridg-*

ing. These two methods were selected from among the 16 maneuvers presented in the BEND framework discussed above.

B.4.1 Exploring Emergent Behavior when Bots *Back* Influencers

Backing involves following and retweeting influencers to amplify their message. *Back* is defined as "actions that increase the importance of the opinion leader" [34]. These actions can be as simple as following and retweeting the opinion leader. In our experimental design we compare the difference between bots randomly retweeting and following versus targeted backing of the influencers (agents with high in-degree).

Figure B.1 visualizes network topology and belief density before and after 500 time steps ($\sim$ 2.5 weeks). From this visualization we see the natural clustering that tends to occur due to homophily, as well as a limited belief uniformity within clusters. In Figure B.1(c) we also see the mean belief fluctuates as bots promote disinformation and *stiflers* try to suppress the disinformation campaign. Tweets by type are provided in Figure B.1(d).

We see that bots initially attract other bots to follow them, and in so doing will slowly get regular users to follow them. At the end of 500 time steps, most bots had 1 or at most 2 normal users following them (as well as 5 to 10 other bots). At face value this mimics what we observe and expect in our empirical bot research.

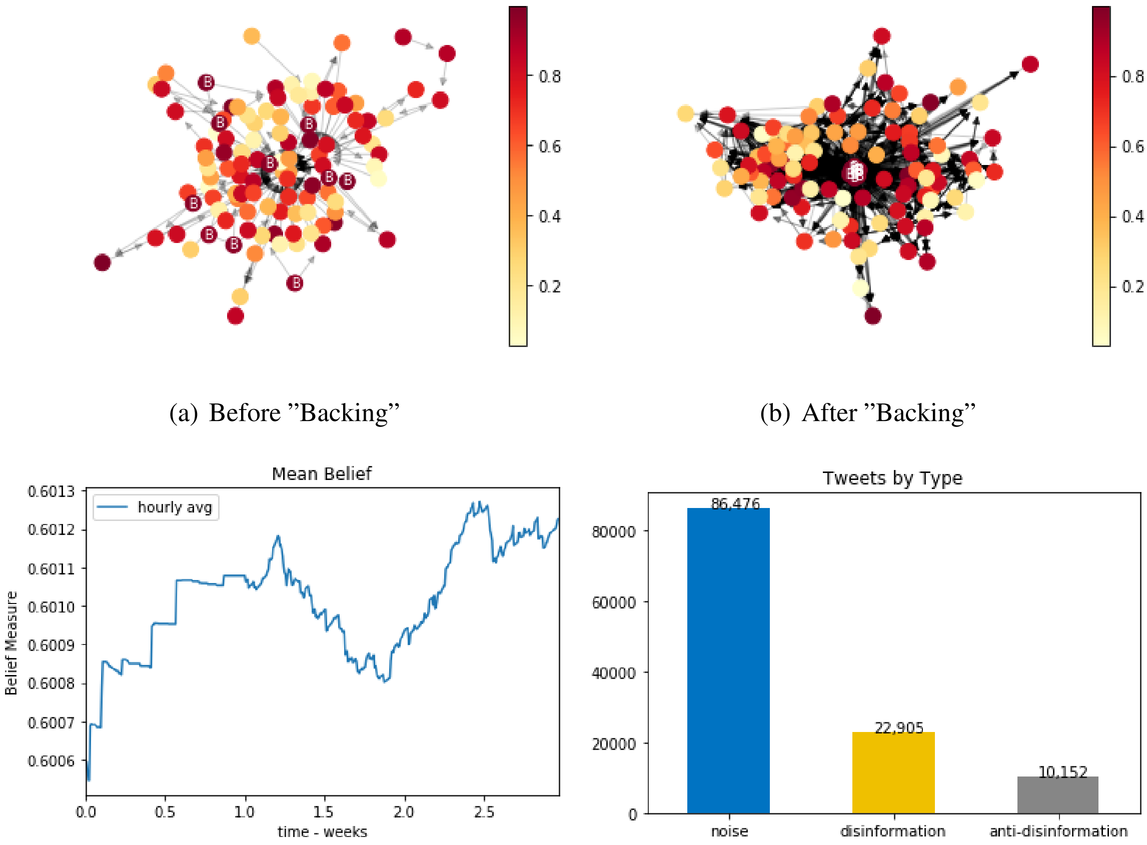
Given that, at face value, this small network mirrors the behavior we expect from bots that *back* and promote, we ran an experiment on larger networks. We conducted a total of 96 runs of the experiment on 1000 node small world networks with bot percentages that range from 5% up to 20%. The experimental design is provided in Table B.3.

Table B.3: Experimental Design for Backing and Bridging

	Bot Percentage					
	5%	8%	10%	12%	15%	20%
Baseline (random following)	12	12	12	12	12	12
Back (targeted following of influencers)	12	12	12	12	12	12
Bridge with Random Following	12	12	12	12	12	12
Bridge with Targeted Following	12	12	12	12	12	12

The results are presented in Figure B.2. From the bottom part of this figure we see that *stiflers* can maintain a downward trend on belief in disinformation until the bot percentage exceeds 12%. At a bot percentage of 12% the battle for belief is generally at a stalemate, and at bot percentages greater than 12% the bot campaign begins to build increasing belief in disinformation.

From the top part of Figure B.2 we see that random following appears to perform better than targeted following in most scenarios, but never with statistically significant results. We found that in the targeted scenario bots were able to leverage the influence and prestige of the influencers to enhance their campaign, but they can't embed in networks and gain followers as fast as the random following scenario. Because the bots that randomly follow were able to embed in local networks and gain followers faster, they achieved at least parity with the bots that target influencers.



(a) Before ”Backing”

(b) After ”Backing”

(c) Mean Beliefs

(d) Types of Tweets Sent

Figure B.1: 100 Node Scale Free Network Before and After 2.5 weeks of Bot *backing* operation and other normal Twitter behavior and network evolution (Bots are labeled with “B”

B.4.2 Exploring Emergent Behavior when Bots ”Bridge” Communities

Bridging involves building links between two communities in order to introduce the ideas of one into the other. *Bridging* is defined as “actions that build a connection between two or more groups” [34]. Our team has observed this behavior in political conversations in both the United States and in Europe. The perpetrators typically identify a target group or community that they want to influence with the ideas and norms of a separate group or community. For example, in US political events our team has observed efforts to connect far-right groups with Christian and other religious communities. The perpetrators first attempt to embed into the target community. Once embedded, they will begin to introduce ideas and create connections between the two groups.

It is important to note that the *backing* algorithms was not changed for *bridging*. The only difference is that the bots were given two communities instead of one. In the process we found that if bots conduct either random or targeted following (or *backing*) when given two communities, they will inevitably “bridge” those communities. This is an important confirmation for us. Even when researchers observe bots with high betweenness and make assumptions of an underlying intention to bridge, it may just be that a bot that is backing influencers has been intentionally or

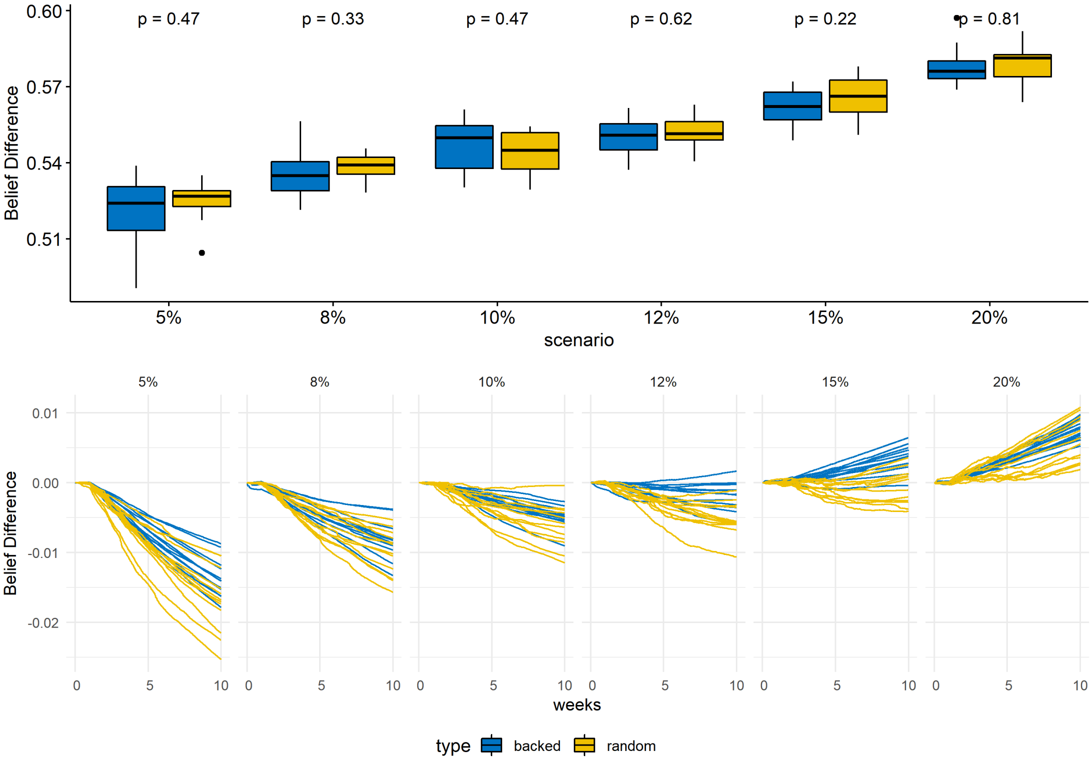


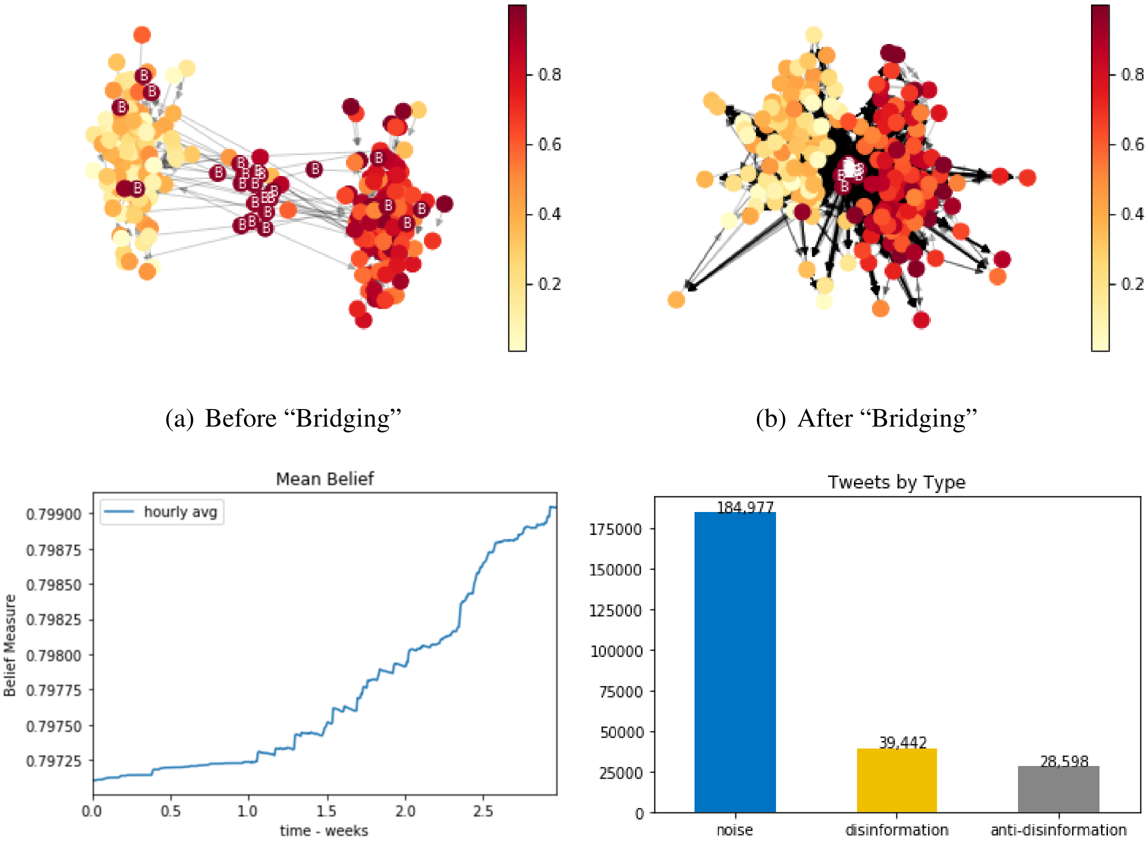
Figure B.2: Results of Backing on 96 runs with random and targeted following

unintentionally oriented on two or more communities.

Figure B.3 visualizes network topology and belief density before and after 500 time steps ($\sim$ 2.5 weeks). Here we see two separate groups, one that strongly believes in the disinformation message, and one that does not. The bots have been inserted in the middle with a single following tie to each of the two communities. In Figure B.3(b) we observe the network after 500 time steps. We see that even after only $\sim$ 2.5 weeks, the bots have already started to bring the two communities together and are starting to introduce the ideas of the one group into the other group. In Figure B.3(c) we see that the mean beliefs of the target community are already starting to increase, demonstrating that they are beginning to believe in the disinformation. In Figure B.3(d) we see the types of tweets sent during the $\sim$ 2.5 weeks.

From this visualization we see that *bridging* does bring the two communities together, and that the *siflers* that are present in the target community are not able to prevent an increase in belief in the disinformation message. We also see the network topology evolve, bringing the two groups together, with the bots having high betweenness centrality.

Given that, at face value, this small network mirrors the behavior we expect from bots that *bridge* two communities, we set to run an experiment on larger networks. We conducted a total of 96 runs of the experiment on 2000 node small world networks with bot percentages that range from 5% up to 20%. The target community consists of a 1000 node small world network with initial beliefs distributed between 0 and 0.5, and the “host” network consists of a 1000 node small world network with initial beliefs distributed between 0.5 and 1. The experimental design



(a) Before "Bridging"

(b) After "Bridging"

(c) Mean Beliefs of Target Community

(d) Types of Tweets Sent

Figure B.3: 100 Node Scale Free Network Before and After Bot 500 hours of Bot *bridging* operation and other normal Twitter behavior and network evolution (Bots are labeled with "B").

is provided in Table B.3.

The results are presented in Figure B.4. The bottom part of this figure is modeling the average belief of the target community (not the whole network). At a bot percentage of 12% the battle for belief is generally at a stalemate, and at bot percentages greater than 12% the bot campaign begins to build increasing belief in disinformation.

From the top part of Figure B.2 we see that random and targeted following appear to be the same. This makes sense given the observation made above, and validates that a bot that is programmed to either back influencers or randomly promote, when intentionally or unintentionally pointed at two communities, will automatically begin *bridging* those communities and communicating beliefs and norms between them.

B.5 Validation

Most rumor propagation models measure the spread of the rumor, not the outcome in people's beliefs. This is primarily because it is easy to measure and validate the spread of the information,

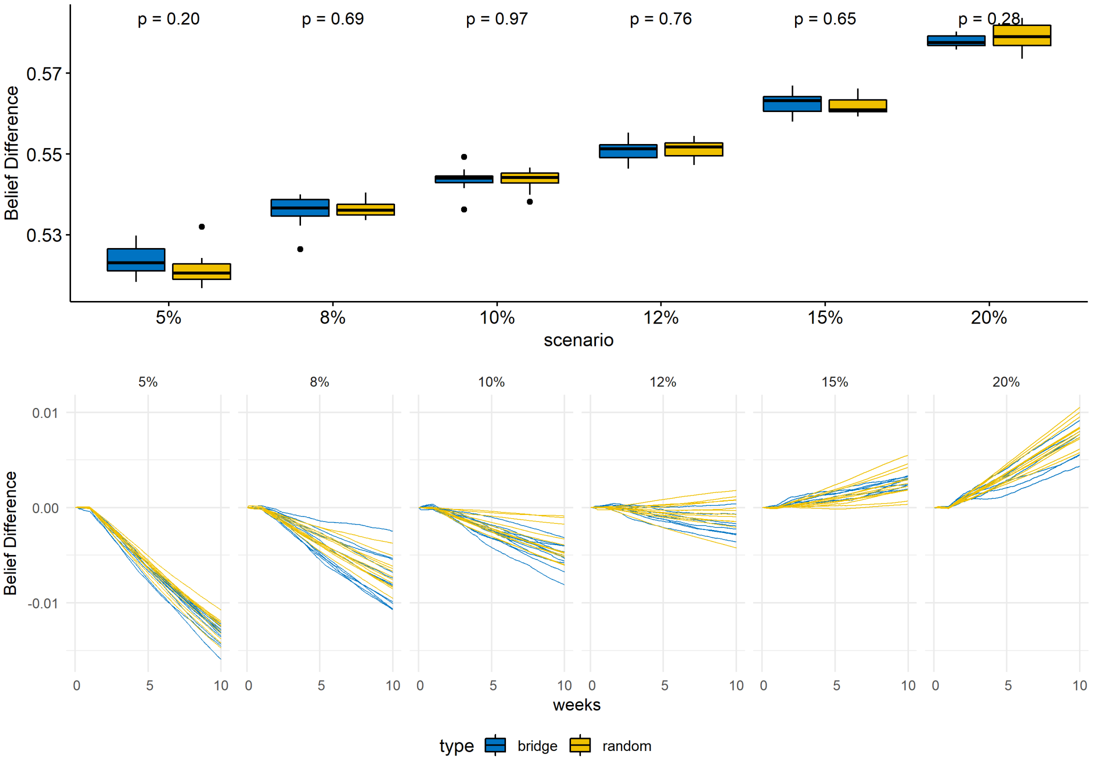


Figure B.4: Results of Bridging on 96 runs with random and targeted following

but is nearly impossible to measure and validate the beliefs of an online community. While proxies may exist in some cases (for example stance can be inferred by some hashtags such as #ClimateHoax or #GunControlNow), these proxies may be weakly correlated and are subject to manipulation (hashtag latching for example). Even though we were not able to exactly model belief, we still kept belief based on exposure as our outcome variable, paired with a simulation that closely replicated the Twitter environment and the actions of agents within this environment.

Given these limitations, we focused our validation efforts on making sure we accurately modeled the behavior that we empirically see in Twitter. The focus of this validation was on estimating the distribution of the inter-arrival time of a Twitter Users' *session* (single log-on episode) and the activity of a single *session*. We also sought to make sure our distribution of original tweets, retweets, and replies replicated the distribution seen in typical Twitter streams.

To validate these metrics we collected three separate Twitter Streams. We collected tweets associated with followers of all US Congressional politicians and congressional candidates of the 2018 mid-term elections. From this we sampled 10M tweets from 56,908 users. The second data set was from tweets associated with users discussing the 2018 Swedish mid-term elections (18M Tweets from 101,260 users). The final data set was a sample of tweets from the 1% Twitter Sample (11M Tweets from 20,144 users). While there is evidence that this 1% sample is not random [175], we still felt that this would give a data perspective that was not necessarily tied to a political event/process.

For each data set we collected the last 200-600 tweets of the users. We then tried to segment the tweet timeline into *sessions*. For the sake of this paper we determined that if any tweet inter-arrival time was greater than 15 minutes, that it constituted a new *session*. If tweet inter-arrival time was less than 15 minutes, we considered it as the same *session*. Having segmented the user timelines, we were then able to calculate tweet inter-arrival time, session inter-arrival time, the user rate (sessions per day), and the tweets per session. Descriptive statistics are provided in Table B.4.

We do want to highlight that numerous bots are present in this data. Using the bot-hunter tool [30] with a threshold of 0.6, the US midterm election data contains 38.9% bots, the Swedish Election data contains 45.7% bots, and the 1% Sample contains 12.8% bots. Some of these bots tweet almost constantly, and therefore all tweets in our sample of their timelines were considered a single *session*. These accounts also can tweet at the speed of algorithms, which is why the minimum inter-arrival time for all three data sets rounds to 0.

Table B.4: Empirical Validation of Twitter User 'Sessions'

Twitter 1% Sample						
	Mean	Min	1st QT	Median	3rd QT	Max
Tweet Interarrival (hrs)	7.15	0	0.011	0.08	1.39	6.82 yrs
Session Interarrival (hrs)	16.5	0.167	0.68	2.1	10.6	6.82 yrs
User Rate (Sessions/Day)	9.73	0.00027	0.0086	0.032	0.15	111K
Tweets per Session	2.3	1	1	1	2	> 3200
Swedish Election Data						
Tweet Interarrival (hrs)	33.5	0	0.027	0.3	10.3	10 yrs
Session Interarrival (hrs)	60	0.167	1.14	7.31	27.3	10 yrs
User Rate (Sessions/Day)	41.2	0.00024	0.004	0.02	0.166	1036K
Tweets per Session	1.76	1	1	1	3	> 3200
US 2018 Midterm Election Data						
Tweet Interarrival (hrs)	61.9	0	0.041	1.27	22	9.14 yrs
Session Interarrival (hrs)	97.8	0.167	1.82	13.3	46.6	9.14 yrs
User Rate (Sessions/Day)	3.7	0.00024	0.0018	0.0071	0.0447	20K
Tweets per Session	1.57	1	1	1	1	> 3200

Given at least one mention, the mean mentions for Sample and Sweden were approximately 1.5 mentions, while US elections was 3 mentions. Note that for all three data sources, the 75th percentile was still 1 mention, indicating a highly skewed distribution.

The number of tweets that are retweets, replies, or contain a mention is provided in Table B.5. From this we see that 40-50% of tweets are tweets, and 20-25% are a reply. We also see that approximately 70-80% contain a mention (note that all retweets contain a mention to the originator).

Table B.5: Portion of Tweets by Type

	Contain Mentions	Is a Retweet	Is a Reply
Twitter 1% Sample	68.80%	48.90%	19.70%
Swedish Elections	80.20%	49.70%	26.50%
US Midterm Elections	67.90%	39.90%	19.40%

B.6 Conclusion

In this paper we've presented the *twitter_sim* ABM designed for exploring the explicit actions users make in Twitter and capturing the varied actions of malicious agents like bots/tolls. We've demonstrated the use of this model in exploring the emerging behavior of specific disinformation maneuvers. Finally, we've validated some of the key variables in the model from empirical Twitter data.

From use of the *twitter_sim* model in exploring *backing* and *bridging* campaigns demonstrated that bots are not as effective if they don't embed in networks and gain a following, even if they are amplifying the messages of influencers. We also learned from the *bridging* experiment that *bridging* can occur simply by pointing bots that are programmed to *back* at multiple communities. In the process of *backing* key individuals, they will *bridge* the network.

twitter_sim is an extremely adaptable and scalable agent based model that fills some key voids for those studying disinformation, information diffusion, or even marketing more generally. It also offers promise for those seeking a model to create a virtual social media environment for training environments.

Appendix C

Labeling Data with Random String Detection

This chapter offers one approach to annotating data. In this case, we develop a Tier 0 model that can find accounts that have 15 digit randomly generated alpha-numeric strings for their screen name. We have found hundreds of these involved in suspicious activity, and have not found one that clearly has human behavior...these are almost surely all bots. We then deploy the Tier 0 model to label a large number of accounts that we can then use for training our Tier 1, 2, and 3 models on. This approach gives us a large and diverse data set for training.

C.1 Related Work

Classifying strings as *random* or *not random* in order to filter or flag anomalous events has a limited background.

Several methods have been proposed for identifying or highlighting the randomness of character strings. Some have proposed leveraging Shannon's Entropy calculation [209] as a method for sorting strings by a measure of randomness. Some cyber security research teams have proposed a similar detection methods in order to detect domain names that are generated by Domain Generation Algorithms (DGA). These teams have separately used Kullback-Leibler Divergence [258], a dictionary approach [178] and Markov modeling [196].

The past research most closely connected to our effort was conducted by LinkedIn in 2013. At that time [99] presented the application of the Naive Bayes model on Character N-grams features of LinkedIn account names in order to identify *spammy* accounts (first and last name as provided by the account owner). This effort was very effective, and replaced the legacy spam detection models that LinkedIn was using on their OSN. To date, our team has not found any team that has replicated a similar approach to Twitter screen names.

C.1.1 Project background

Our team has focused on detecting, characterizing, and modeling the behavior of bots, bot networks and their creators. In doing this we've studied several recorded bot events. Recently we focused on a known and publicized bot attack against the Atlantic Council Digital Forensic Labs (DFR Lab), and tangentially against the NATO Public Affairs Office. This attack primarily occurred between August 28 and August 30, 2017. We also focused on a recorded bot harassment event against journalists in Yemen [159]. In both events we observed numerous bot accounts that used 15 character randomly generated alpha-numeric strings for the screen name. Examples of this include **Wy3wU4HegLlvHgC**, **5JSQavWW3tvQwA7**, and **gG6RKc6QBqOLKyU** (these are not real Twitter accounts). Note that these randomly generated strings always sample from upper and lower case alpha-numeric characters. Observing this phenomenon motivated the construction of this algorithm and its application on Twitter at large in order to observe other bots and bot actors that are using these same types of bot screen names. More importantly, we hope this dataset can be used as a large and diverse annotated bot training data for larger and more comprehensive machine learning models.

C.2 Modeling

C.2.1 Feature engineering

In order to develop a random string detection model for this unique case, we constructed training data consisting of 200,000 non-random Twitter screen names (randomly sampled from Twitter and manually verified as non-random) and 200,000 randomly generated 15 digit strings. We then developed a combination of heuristic filtering and traditional machine learning models to label the string as *random* or *not random*. This development is described below.

For feature engineering, the primary feature that we extracted from the strings was character n-gram. For string s with length m, a character n-gram is the $(m - n + 1)$ sequential substrings of length n found in string s. In our case, we explored several settings for n, to include using multiple values in the same feature set (i.e. using both bigrams and trigrams).

We then transformed the resulting sparse character n-gram matrix using term frequency-inverse document frequency (TF-IDF). TF-IDF is defined in Equation 1 and 2 below, and is used to scale the characters by the information that they provide. In our case, frequent characters in a string provide information, but not if they're frequent in all of the strings. To calculate the IDF for character c in strings s, we take the logarithm of the ratio of the total number of strings in corpus S by the number strings that contain c, as shown in Equation C.1.

$$idf(c, S) = log \frac{N}{|\{c \in S : c \in s\}|} \tag{C.1}$$

We then calculate the TF-IDF for character c in string s found in corpus S as follows

$$tfidf(c, s, S) = tf(c, s)\dot{i}df(c, S) \tag{C.2}$$

This therefore weights characters that have a high local frequency but a lower global frequency. At first it may seem that TF-IDF is unnecessary since each character n-gram is equally likely in random strings, given a strong pseudo-random number generator. n-grams are not equally likely for human generated strings, however. Given this fact we felt it appropriate to transform the data with TF-IDF.

These features were merged with several other features. We started by merging the normalized count of upper case, lower case, and numeric characters. n-gram generation by default converts all text to lower case. We maintained this default behavior, but saw that the number of upper and lower case in letters in particular provided a strong signal. Since our training data contained some human generated strings that were not 15 characters in length, we normalized these counts.

Additionally, we included the Shannon string entropy in our feature set. Shannon string entropy, while not strong enough to use by itself in our case, still provides a strong signal that we felt would be useful. We will test this assumption below. Shannon entropy is defined in C.3, where p_i is the normalized count for each character found in the string.

$$H\left(A\right) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{C.3}$$

The A full table of features is given in Table C.1.

Table C.1: Features for Random String Detection

Feature	Type	Description
Character Bi-gram	Numeric	Term frequency inverse document frequency of bi-gram
No. lower case	Numeric	Normalized count of lower case letters
No. upper case	Numeric	Normalized count of upper case letters
String entropy	Numeric	Shannon String entropy

We used the $scikit-learn$ package [189] to explore and build the machine learning classification model for Random Strings. We evaluated Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) with 10 fold cross-validation. The results are presented in Table C.2. We conducted model comparisons between these models, and found SVM and Logistic Regression did are not statistically different ($t = 0.62912$, $df = 18$, $p.value < 0.5372$). Given these results, we used Logistic Regression for our production model, given that it is simpler and faster. Note that this result entails significantly more training data than we used in earlier research (see [31]), where SVM performed better.

Table C.2: Model Performance in Classifying Randomly Generated Strings for Screen-names

Model	Accuracy	F1	Kappa	Precision	Recall	ROC AUC
Log. Regression	0.996	0.996	0.991	0.994	0.997	0.999
Naïve Bayes	0.969	0.97	0.939	0.947	0.995	0.996
SVM	0.996	0.996	0.993	0.995	0.998	1

Before predicting whether a string was random, we first applied several heuristic filters. These verified that 1) the string was 15 characters in length, and 2) contained at least one capital letter, lower case letter, and numeric digit. This final filter was applied given that 15 character strings have a 0.02% chance of not containing a capital or lower case letter and a 7% chance of not containing a numeric digit. This heuristic was applied given that precision was a higher priority than recall.

In Figure C.1 we evaluate the best value of n (number of characters for n-gram) as well as whether or not using Shannon's Entropy as a column feature provides leverage in prediction. In this visualization we see that bigrams with Shannon's entropy provides the best leverage in predicting random strings.

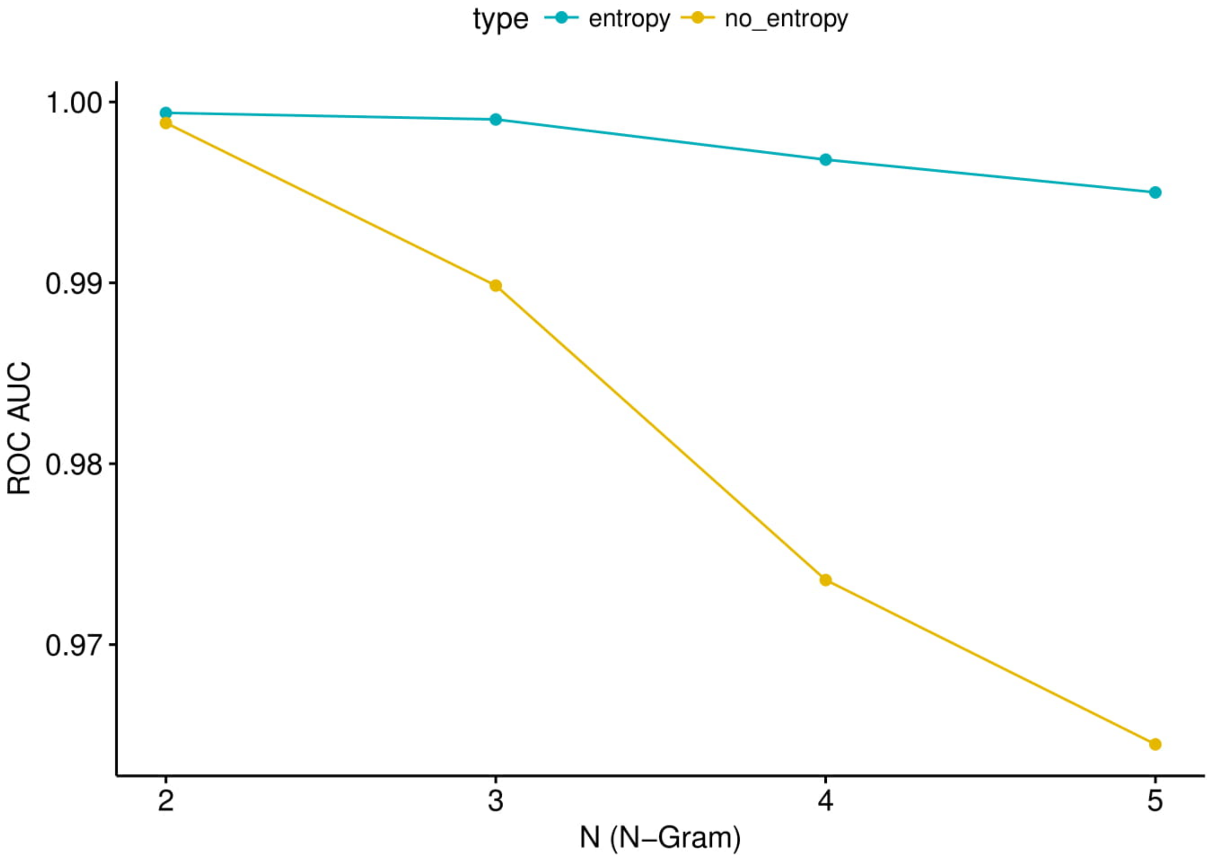


Figure C.1: Evaluating n (number of characters in n-gram) and use of Shannon's entropy as a feature

In addition to exploring the feature-based machine learning models discussed above, we also explored the use of Markov model of character sequencing, but found during initial exploration that this did not have sufficient power to classify the strings given the inherent random nature of human generated screen names. Additionally, we explored using Shannon entropy as the only measure for filtering these strings. Once again, while helpful, this method did not demonstrate sufficient power for our purposes.

C.2.2 Model Deployment

Our primary use for the algorithm was to filter accounts with 15 character random strings from a Twitter data stream. To do this we ran a random sample from the Twitter Streaming API from 23 December 2017 to 20 June 2018. During this time the stream collected approximately 433 million tweets. This collection was done without any semantic or geographic filters, and we stored the raw JSON files that are returned by the Twitter API.

Having performed the collection, we next applied our algorithm to all 433 million tweets, filtering out all accounts that were labeled as having 15 digit randomly generated screen name. This produced a collection of 7.8 million tweets from 1.7 million unique accounts.

C.3 Model Evaluation

Given the desired use case of annotating diverse bot accounts, we conducted two evaluations on our results. First, we wanted to estimate the false positive rate on our random string detection, since false positives have a high likelihood of not being an autonomous account. To accomplish this we randomly selected 1,000 of the screen names that were labeled as random, and manually identified those that contained clear words or acronyms. Given this method, we estimate that our false positive rate is approximately 1%.

Additionally, we wanted to estimate the percentage of random character screen name accounts that are automated, or appear automated. In other words, how many of our true positive random string accounts are truly bots. To estimate this, we randomly sampled 100 accounts, verified that the user name appeared random, and inspected the account in the Twitter web client. Of the 100 that we manually inspected, five were suspended, eight provided no results (most likely the account was closed by the user), and all others exhibited autonomous behavior. After thoroughly evaluating these 100 randomly sampled accounts we were satisfied that this methodology provides annotated bot data that is at least as accurate as *honey pot* data, and likely has a wider range of bot types.

C.3.1 Data Characterization

One of our first tasks in exploring the data is to understand how these accounts differ from the average Twitter account, and whether those differences were uniform across the language of the bot creator.

99% of the 7.8 million tweets in this dataset are associated with seven languages. It's interesting to note that none of the Continental European Languages (French, Spanish, German, Portuguese, Italian, etc) are in this list. Somewhat surprisingly, the proportion associated with Japanese and Arabic accounts is very high, second only to English. A full breakdown of the languages and a short general description of our observations are provided in Table C.3. Only 840 tweets contained coordinate locations, and these locations are strongly correlated to the languages mentioned below (United States, Japan, the broader Middle East, Russia, and Thailand).

The major observations from Table C.3 are that the random string accounts are younger, less popular, and less active than the average Twitter account. We see that the median age for the random bots is 224 days, compared to 1,248 days for your average active Twitter account. The median number of followers/friends ratio for the random string bots is 6/39 versus 277/294 for the average Twitter account. We also see that the median random string bot account only produced 54 tweets over its lifetime, versus 8,216 for the average account (this comparison is affected by age difference).

Table C.3: Summary Statistics by Language

Language		Arabic	English	Japanese	Korean	Russian	Thai	Chinese	other	Normal*
# of Accounts		246K	626K	593K	103K	61K	47K	21K	18K	1599K
Age	min	61	61	61	61	62	61	62	61	6
	25%	181	105	214	193	162	167	186	192	487
	50%	264	165	361	260	292	246	288	297	1,248
	75%	413	213	570	427	365	383	423	626	2,235
	max	3,046	17,763	3,731	3,020	3,075	3,306	3,431	3,662	4,421
	mean	326	210	449	342	322	310	357	550	1,412
	std	216	253	315	229	228	219	247	609	1,008
Followers Count	min	0	0	0	0	0	0	0	0	0
	25%	3	0	2	0	0	1	0	2	78
	50%	15	2	19	2	1	5	1	15	277
	75%	63	17	108	10	5	24	6	85	818
	max	828K	1087K	1322K	23,681	54K	177K	50K	944K	40,550K
	mean	171	61	136	32	93	163	97	295	3376
	std	2,716	2,054	2,366	268	1,013	2,044	921	7,581	94990
Friends Count	min	0	0	0	0	0	0	0	0	0
	25%	26	5	10	1	6	31	7	29	118
	50%	79	26	49	21	31	88	32	91	294
	75%	226	73	168	74	53	258	70	242	695
	max	640K	349K	75K	25K	17K	18K	12K	88K	2,441K
	mean	297	101	178	92	130	257	98	298	1044
	std	1,543	682	482	326	594	498	322	1,034	8227
Tweet count	min	1	1	1	1	1	1	1	1	1
	25%	15	6	25	26	16	24	15	19	1813
	50%	71	20	117	134	69	99	59	109	8216
	75%	319	83	515	601	234	370	286	627	27318
	max	532K	806K	994K	228K	114K	570K	106K	304K	16,176K
	mean	819	301	930	934	456	684	517	1,727	26,652
	std	3,753	3,180	4,301	3,652	2,226	4,195	2,036	7,981	66,180

* *Normal* Twitter Accounts were sampled from the Twitter Streaming API

While some languages (Arabic, Japanese, Korean, and Thai) appear to be slightly more popular and active, in general these random string accounts appear to have a high number of accounts that are dormant, or at least in a state of low activity. Some of these may be waiting to be activated for a given event or task, while others may be used for intimidation attacks (as some of these were with the Yemen journalist discussed above). Intimidation accounts (accounts that follow a user in mass) do not need to be active or popular. Their intent is to push another account out of the Twitter conversation through intimidation.

Given the fact that our data set contains primarily bot accounts, we observed a number of account suspensions during the course of our study. Between mid-December 2017 and August 22 2018, 247,022 accounts (~15%) were suspended by Twitter, while 46,985 accounts (~2.7%) were removed by the user. As the media and politicians put pressure on Social Media companies, the natural response is to increase their policing of this automated behavior on their platforms.

Appendix D

Model Execution Times

This appendix provides computational time to execute the models presented in this thesis. The times listed in Table D.1 were all conducted on Twitter JSON data and include the overhead of reading and processing the Twitter data. All tests were executed on a server running the Ubuntu 18.04 operating system, an Intel(R) Core(TM) i7-6900K CPU (3.20GHz) processor with 16 cores, 128GB or RAM and 4 Titan Xp GPUs. Execution time (Wall Time) is reported. All models were run on a single thread (though dependent packages may use multi-threading).

Table D.1: Measuring Execution Time for Models

Model Family	Model Name	Computational Task	Wall Time (hh:mm:ss)
Bot-Hunter	Tier 0	Classify 1000 Twitter Accounts	00:00:00.4
	Tier 1	Classify 1000 Twitter Accounts	00:00:02
	Tier 2	Classify 1000 Twitter Accounts	00:15:43
	Tier 3	Classify 100 Twitter Accounts	08:41:10
	Graph-Hist Tier 3	Classify 100 Twitter Accounts	04:31:00
External Model	Botometer	Classify 1000 Twitter Accounts	02:50:00
Meme-Hunter	OCR	Extract text from 1000 images	00:10:07
	Vision Only	Classify 1000 images	00:01:29
	Multi-Modal	Classify 1000 images	00:22:55
	Template Based	Classify 1000 images	00:01:07
Bot-Match	Cosine Similarity	Create Document-Term Matrix for 1K Accounts	00:00:31
		Create Document-Term Matrix for 10K Accounts	00:14:41
	LDA	Embed Content of 1K Accounts	00:05:45
		Embed Content of 10K Accounts	00:36:52
	Node2Vec	Embed Twitter Network of 1K Accounts	00:04:58
		Embed Twitter Network of 10K Accounts	00:33:10
	Facebook BigGraph	Embed Twitter Network of 1K Accounts	00:01:02
		Embed Twitter Network of 10K Accounts	00:07:39
	GCN (10 epochs)	Embed Twitter Network of 1K Accounts	00:13:06
		Embed Twitter Network of 10K Accounts	00:41:30
Bot-Labels	Normal	Characterize 100 Twitter Accounts	01:36:39
	Fast	Characterize 100 Twitter Accounts	00:13:25

Appendix E

Bot Field Guide

> No effort to thwart modern disinformation efforts will be successful without educating the citizens of open democratic countries of the threats that exist in the modern social media environment. Bots, trolls, cyborgs, sockpuppets, and other agents are increasingly used to manipulate the marketplace of beliefs and ideas in social media, and the average social media user must learn to distinguish these accounts. This bot *field guide* is designed to help social media users understand and recognize these accounts. It is also used by researchers to help understand differences in statistics as well as differences in the performance of bot detection algorithms.

Although social media bots can create positive effects, a subset of malicious bots have recently gained widespread notoriety for their intervention and manipulation of the marketplace of information, ideas, and belief. These bots have been documented manipulating various world events, ranging from manipulating democratic election events to intimidating journalists and researchers. This new threat has led to an emerging discipline often called social cyber security. Social cybersecurity is focused on defending our citizens from malicious elements using technical means to hack our society. While the means are technical, the target is very human.

Many policy recommendations have been made to counter this emerging threat. Many of these policy recommendations are made without knowledge of the underlying technology, and will either be useless or possibly counter productive. For example, some recommend requiring the social media companies to provide users a means to flag "fake news", thereby crowd sourcing the effort. This recommendations fails to realize that if social media companies expose this functionality on their API or just on the front end of their application, then the same bot puppet masters now have a way to flag all content as "fake" at the speed of algorithms. This would make truth even more elusive than it is already today.

While many of the proposed policies are questionable at best, the only recommendation that is whole-heartedly endorsed by all parties is the requirement to educate the citizens of open democracies about how to vet news and sources in the modern information environment.

Toward this end, we have created a bot *field guide* to help internet users see and explore some of the malicious agents that are present on social media platforms, particularly the Twitter environment. While this *field guide* does not cover all possible bot types and actions, it does provide a diverse sampling of the types of bots that exist and that our team in the CASOS lab at Carnegie Mellon University have found in disinformation operations.

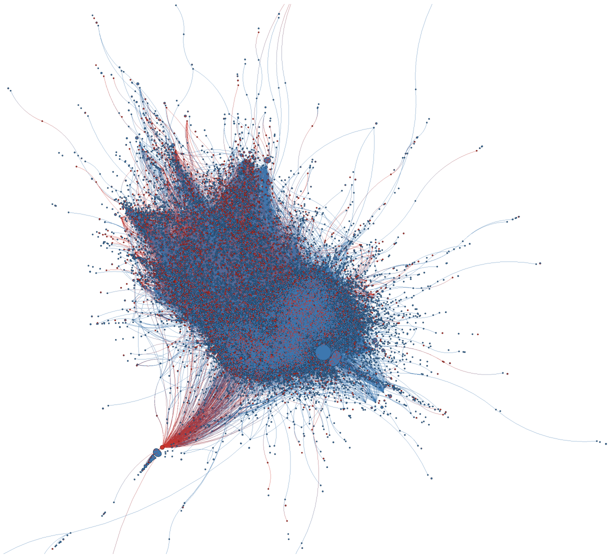


Figure E.1: Bots (red) involved in manipulating the conversation surrounding the Swedish national elections in 2018

For each account we provide a short summary, some visualization of timing, content, and followers, and finally some metrics regarding the account and what several modern bot detection algorithms predict regarding whether or not the account is a bot.

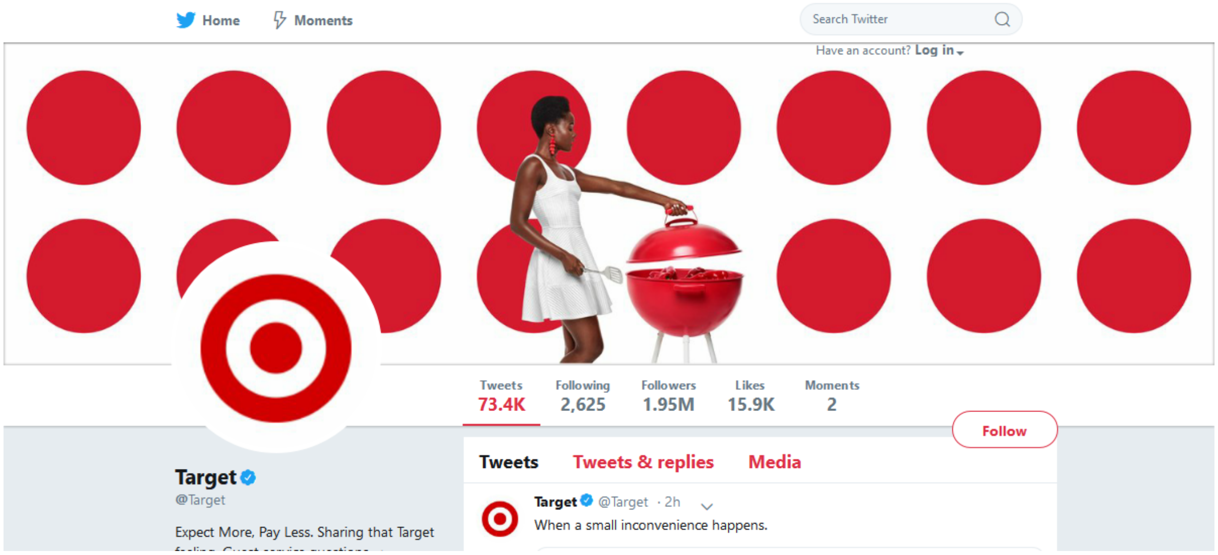
The *field guide* has the following sections:

1. NormalUsers (Personal, Commercial, Government)
2. Amplifier bots
3. Cyborg Bots
4. Chaos Bots
5. Coordinated Bots
6. Social Influence Bots
7. News Bots
8. Overt Bots
9. Intimidation Bots
10. Scam Bots
11. Bots related to specific countries
12. Random String Bots

E.1 Personal, commercial, and government accounts

Before we look at bot, troll, and cyborg accounts, lets first look at a few examples of normal accounts. It helps to have a better understanding of what “normal” looks like before we look at anomalies.

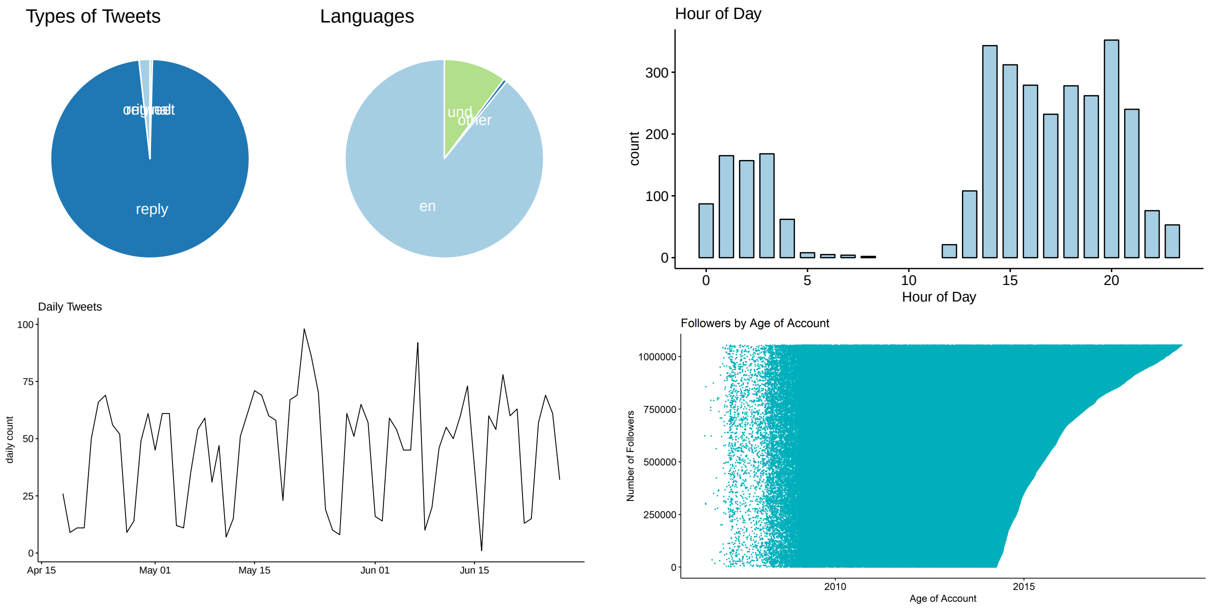
The following several pages will look at several commercial and government accounts. Take a look at the statistics and patterns of these “normal accounts” so that you can learn to distinguish these from anomalous accounts.



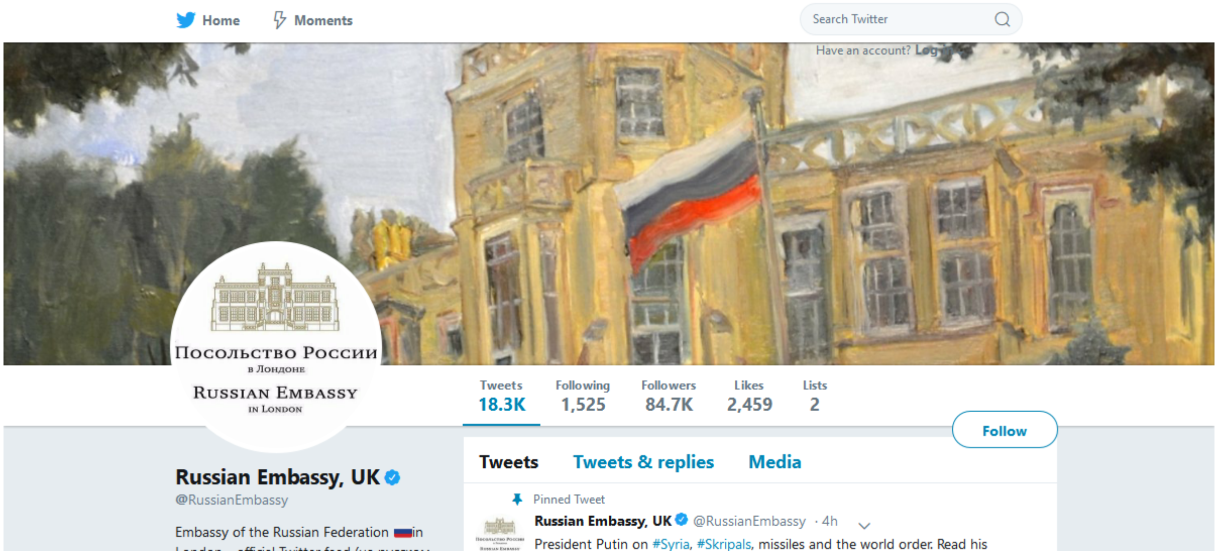
Expect More, Pay Less. Sharing that Target

When a small inconvenience happens.

A. @target. We added Target to give an example of a commercial entity. Note that most of the activity is replies as the Target PR team works to interact with customers and maintain a strong brand reputation.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.001	bot-hunter Tier 0	False	TargetRun	69	StealthedArrow	9	tgt.biz	27.0
Avg Sentiment (0-1)	0.733	bot-hunter Tier 1	0.5912	vineyardvinesForTarget	57	NatGotti	5	target.com	2.0
Dormant Followers	51.9%	bot-hunter Tier 2	0.434	TargetHappens	21	TargetNews	4	bit.ly	1.0
Bot Followers	50.0%	bot-hunter Tier 3	0.409	TargetLittle	17	giannacollins33	4		
Avg Mentions	1.021	botometer	0.0656	TakePride	12	katscurious	4		
Retweets RU Prop.	False	Debot	False	TargetRuns	8	vineyardvines	3		



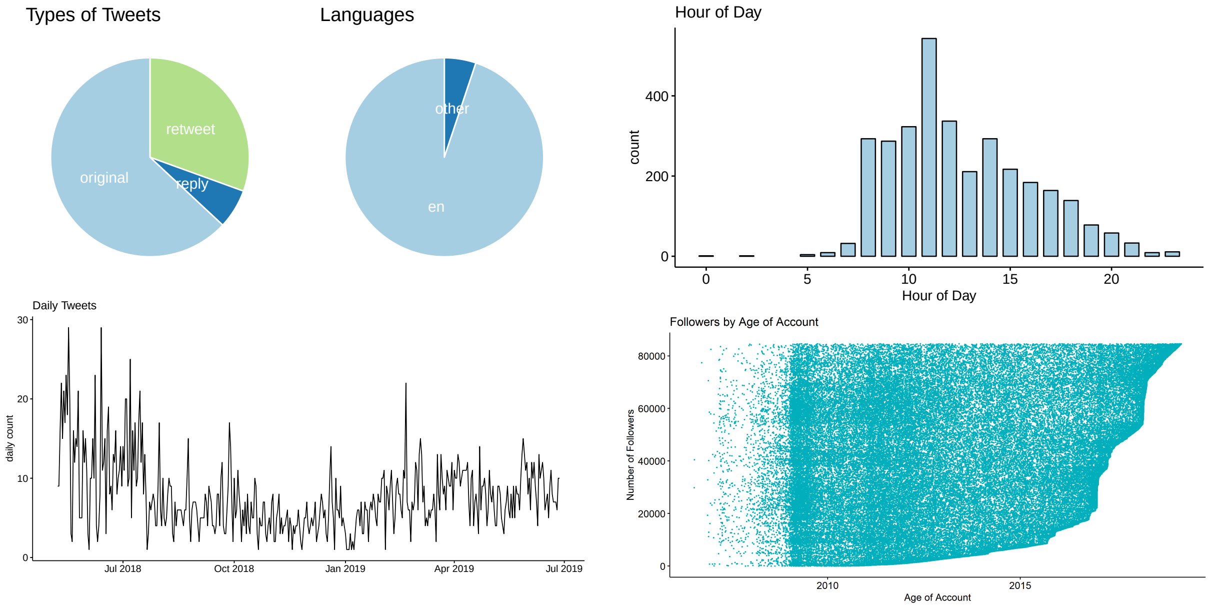
Russian Embassy, UK
@RussianEmbassy

Embassy of the Russian Federation in

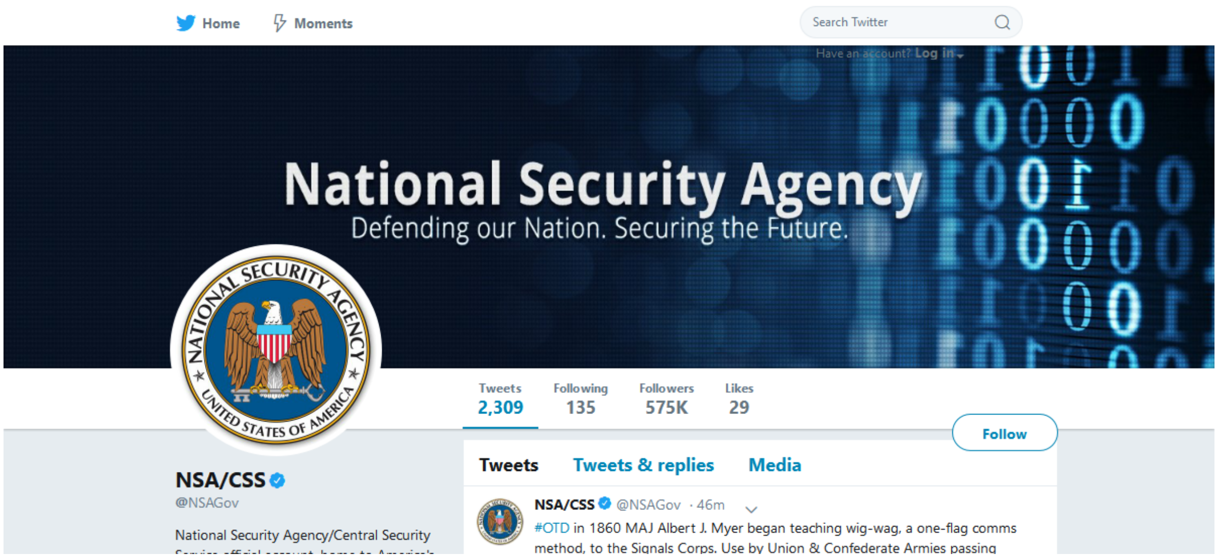
Russian Embassy, UK @RussianEmbassy · 4h

President Putin on #Syria, #Skripals, missiles and the world order. Read his

B. @RussianEmbassy. This shows the activity of the Russian Embassy to the UK. Note that the Russians Embassies play an important role in Russia's disinformation campaigns. Also note that the bot-hunter suite of tools, trained to find Russian bots/trolls, classifies the Russian Embassy as a bot.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.018	bot-hunter Tier 0	False	Lavrov	241	Amb_Yakovenko	433	rusemb.org.uk	71
Avg Sentiment (0-1)	0.686	bot-hunter Tier 1	0.7374	Zakharova	127	mfa_russia	176	bit.ly	13
Dormant Followers	30.7%	bot-hunter Tier 2	0.4982	WorldCup	86	RSGovUK	150	sciencedirect.com	7
Bot Followers	51.2%	bot-hunter Tier 3	0.4220	Russia2018Quiz	82	FIFAWorldCup	79	pscp.tv	7
Avg Mentions	1.379	botometer	0.1033	Salisbury	41	KremlinRussia$_E$	51	rt.com	7
Retweets RU Prop.	True	Debot	False	Russia	41	thetimes	37	mid.ru	6

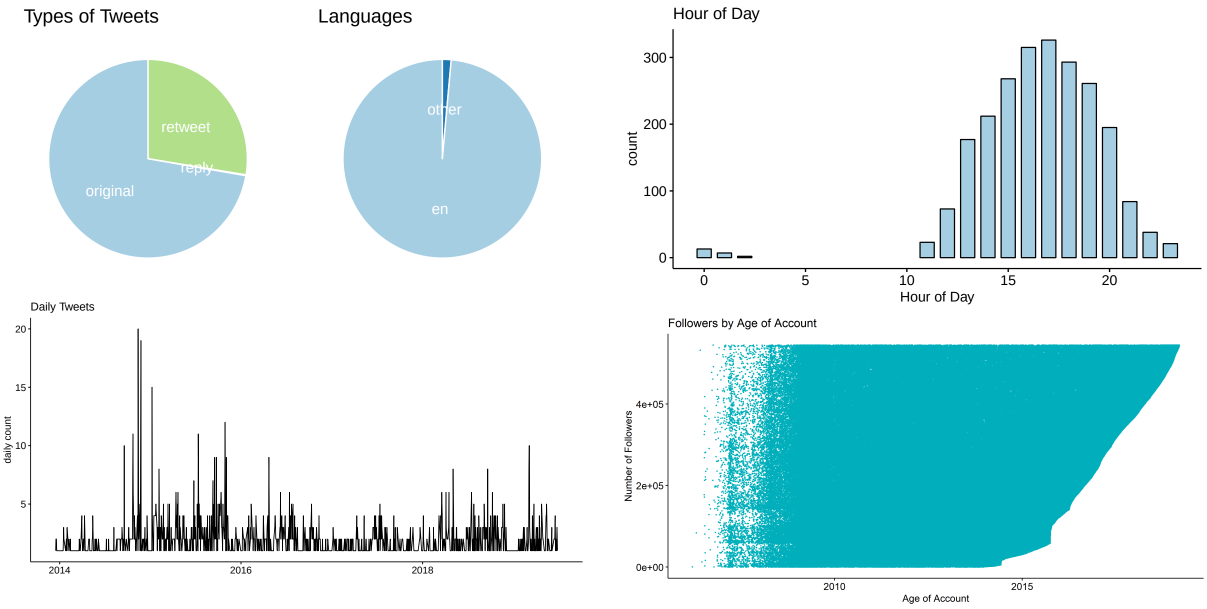


NSA/CSS
@NSAGov

National Security Agency/Central Security

NSA/CSS @NSAGov · 46m
#OTD in 1860 MAJ Albert J. Myer began teaching wig-wag, a one-flag comms method, to the Signals Corps. Use by Union & Confederate Armies passing

C. @NSAGov. This gives an example of a US government agency. Note the account typically produces original content during a consistent 8 hour work day. Note that none of the bot detection methods falsely label this as a bot.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False	NSA	756	NSACareers	187	bit.ly	339
Avg Sentiment (0-1)	0.702	bot-hunter Tier 1	0.3811	news	115	NSAGov	99	usa.gov	324
Dormant Followers	35.2%	bot-hunter Tier 2	0.272	GenCyber	98	ODNIgov	86	ow.ly	168
Bot Followers	52.9%	bot-hunter Tier 3	0.463	CryptoChallenge	96	icontherecord	37	fb.me	25
Avg Mentions	1.34	botometer	0.1196	cybersecurity	71	DeptofDefense	35	tumblr.com	16
Retweets RU Prop.	False	Debot	False	TBT	68	WestPoint_USMA	19	facebook.com	14

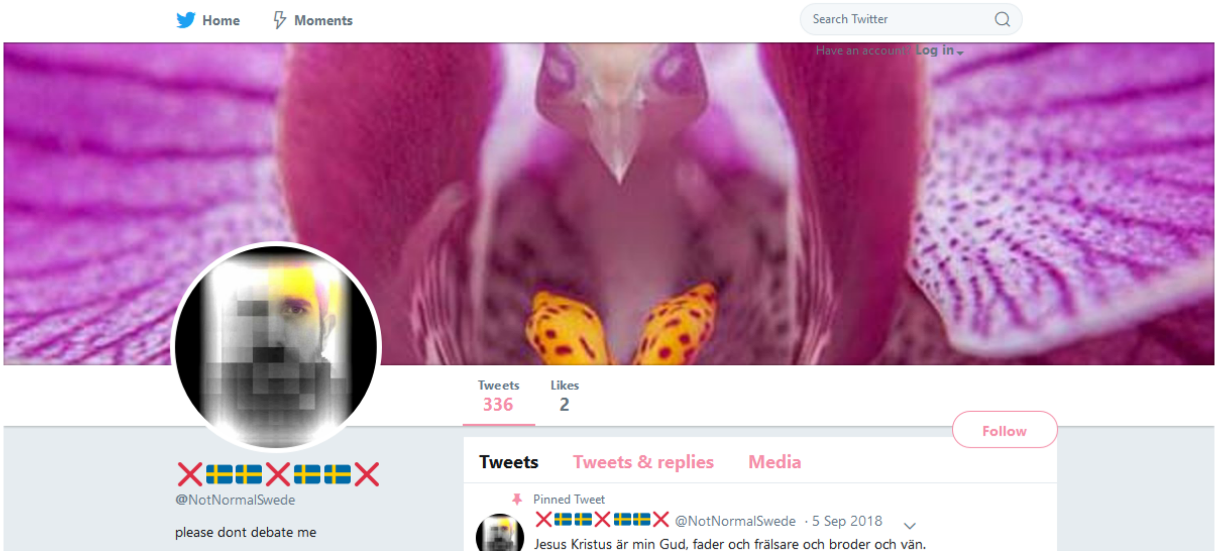
E.2 Amplifier Bots

Amplifier bots are used to propagate a message or amplify the voice of certain individuals/accounts. The easiest way to do this is through retweets and likes. It only takes a few lines of code to create an amplifier bot.

To create an amplifier bot focused on a message, the bot creator could create a bot that retweets any tweet with a given hashtag. For example, a bot can monitor the twitter stream, and every time it detects the hashtag #flatearth, retweet the message.

To create an amplifier bot focused on amplifying an individual, the bot creator would simply create a bot that monitors all tweets produced by the account benefiting from the bot activity, and retweet all tweets.

More sophisticated bots could create more sophisticated rules for choosing what to amplify.

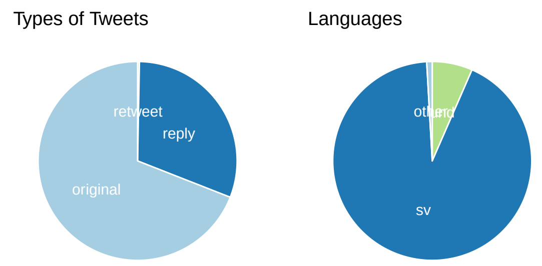


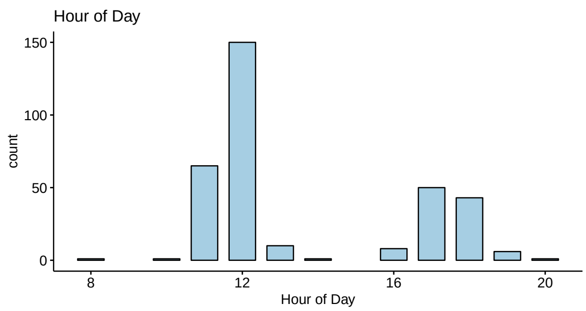
❌🇸🇪🇸🇪❌🇸🇪🇸🇪❌
@NotNormalSwede

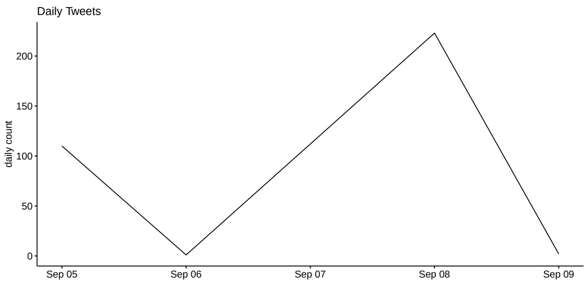
please dont debate me

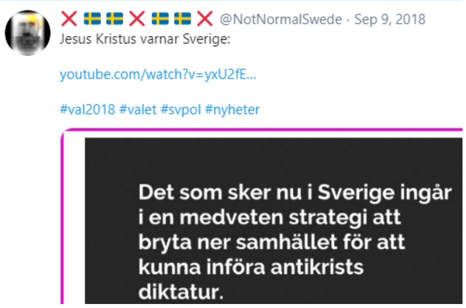
❌🇸🇪🇸🇪❌🇸🇪🇸🇪❌ @NotNormalSwede · 5 Sep 2018
Jesus Kristus är min Gud, fader och frälsare och broder och vän.

D. @NotNormalSwede. @NotNormalSwede was activated just for the Swedish Elections. Notice that it had zero friends or followers. It tweeted with Christian jargon that contained links to YouTube videos that contained strong racist, anti-Semitic and anti-immigrant arguments wrapped in religious argument.









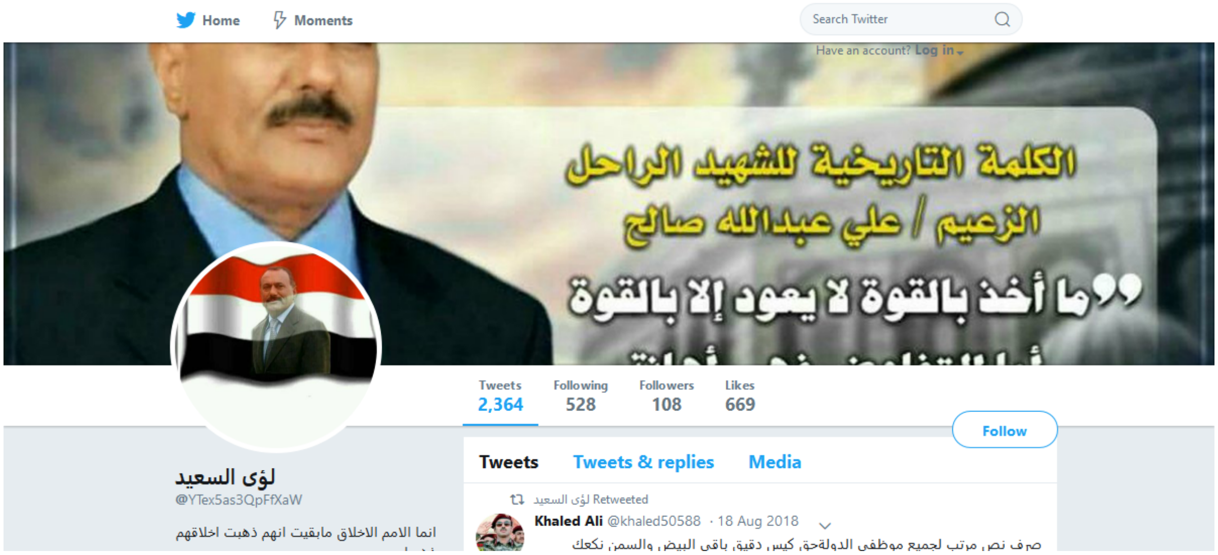
❌🇸🇪🇸🇪❌🇸🇪🇸🇪❌ @NotNormalSwede · Sep 9, 2018
Jesus Kristus varnar Sverige:

youtube.com/watch?v=yxU2fE...

#val2018 #valet #svpol #nyheter

Det som sker nu i Sverige ingår i en medveten strategi att bryta ner samhället för att kunna införa antikrists diktatur.

Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False	SVERIGE	79	SvKyrkansUnga	3	youtube.com	1.0
Avg Sentiment (0-1)	0.6	bot-hunter Tier 1	0.2455	svpol	1	svenskakyrkan	3		
Dormant Followers	NA	bot-hunter Tier 2	0.4175	nyheter	1	zhekaqwet	3		
Bot Followers	NA	bot-hunter Tier 3	0.452	valet	1	UtrotaIslam	2		
Avg Mentions	3.88	botometer	0.7248	minförstatweet	1	buroa_sweden	2		
Retweets RU Prop.	False	Debot	False	val2018	1	Samah08969370	2		


الكلمة التاريخية للشهيد الراحل
الزعيم / علي عبدالله صالح
ما أخذ بالقوة لا يعود إلا بالقوة

Home Moments

Search Twitter

Have an account? Log in

Tweets	Following	Followers	Likes
2,364	528	108	669

Follow

لؤى السعيد
@YTex5as3QpFfXaW

انما الامم الاخلاق مابقيت انهم ذهبت اخلاقهم

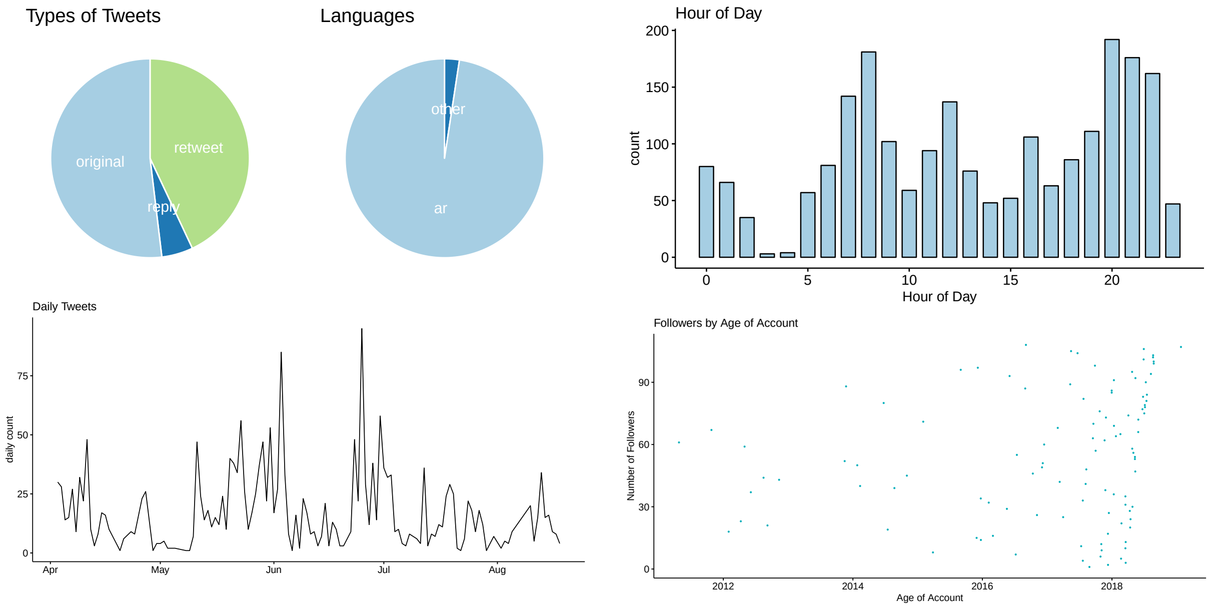
Tweets Tweets & replies Media

لؤى السعيد Retweeted

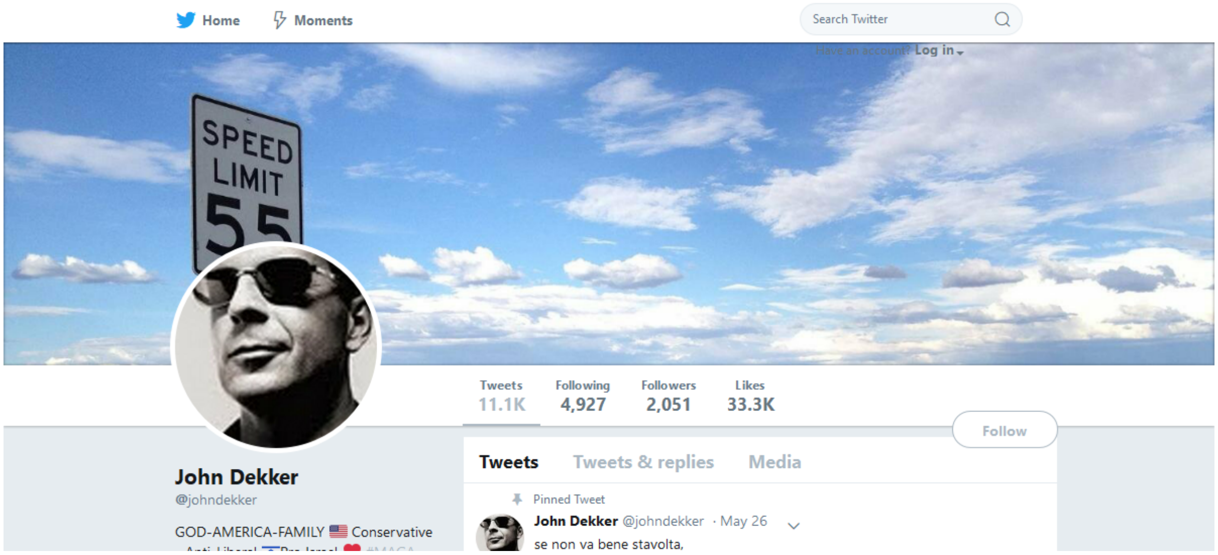
Khaled Ali @khaled50588 · 18 Aug 2018

صرف نص مرتب لجميع موظفى الدولةحق كيس دقيق باقى البيض والسمن نكعك

E. @YTex5as3QpFfXaW. YTex5as3QpFfXaW was discovered in the Yemen data; most retweets appear to be anti-Saudi Arabia and anti-coalition. All original tweets are dumb tweets. Note that all bot models don't predict this as a bot. It maintains a semi-normal tweet rate and ratio of original to retweet to reply. The only way to identify this as a bot is to manually see that all original tweets are dumb tweets and match that with the fact that the account doesn't have a normal circadian rhythm.



Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Domain**	**Count**
Frd/Fol Ratio	5.421	bot-hunter Tier 0	True	*Hodeidah*	17	Ali_Albukhaiti	57	rt.com	2
Avg Sentiment (0-1)	0.671	bot-hunter Tier 1	0.0998	*Houthi*	13	HSaqqaf	27	erem.news	1
Dormant Followers	35.8%	bot-hunter Tier 2	0.112	*Sana'a*	11	akramHajarr	23	facebook.com	1
Bot Followers	18.9%	bot-hunter Tier 3	0.264	*To whom*	11	yawh	22	Wthker.com	1
Avg Mentions	1.048	botometer	0.6729	*Houthis*	9	khaled50588	22	khabaragency.net	1
Retweets Russia Propaganda	True	Debot	False	*The Houthis*	8	MohAlmaswari	21	ghrebaa.com	1



Tweets 11.1K | Following 4,927 | Followers 2,051 | Likes 33.3K

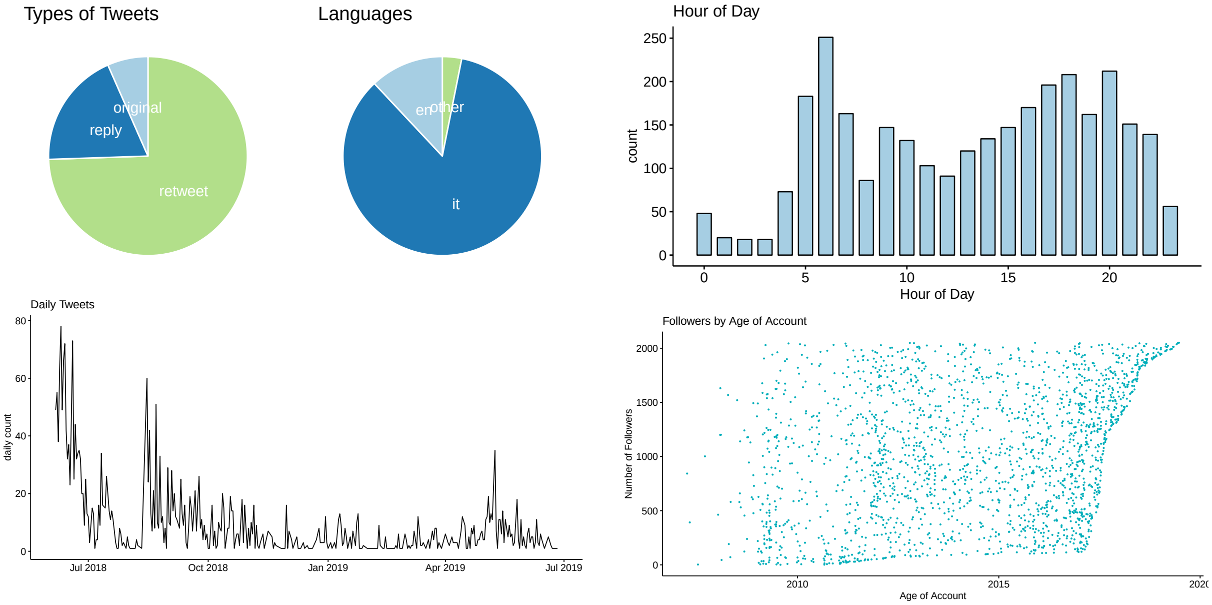
John Dekker
@johndekker

GOD-AMERICA-FAMILY 🇺🇸 Conservative
Anti-Liberal Pro-Israel ❤ #MAGA

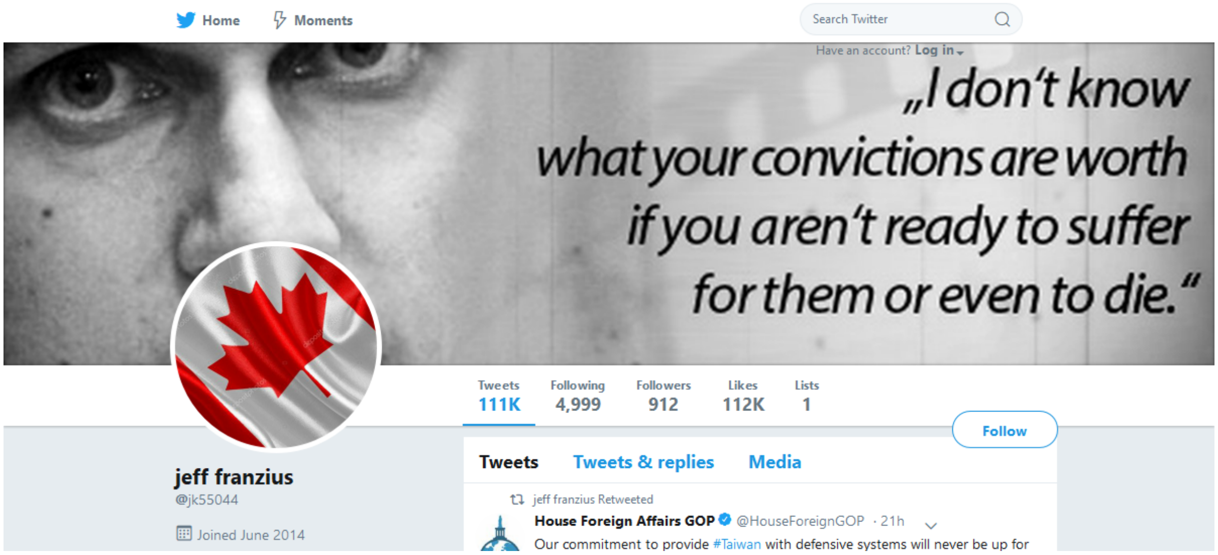
Tweets | Tweets & replies | Media

Pinned Tweet
John Dekker @johndekker · May 26
se non va bene stavolta,

F. @johndekker. We found @johndekker spreading anti-EU disinformation in Italy. John Dekker (with description MAGA, GOD-AMERICAN-FAMILY) seems to live in Italy, posts primarily in Italian, and almost exclusively focused on Italian politics and encourage Italian exit from EU. This shows how bots can be moved around the globe to participate in various conversations.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	2.402	bot-hunter Tier 0	False	Salvini	23	matteosalvinimi	153	fllwrs.com	114
Avg Sentiment (0-1)	0.668	bot-hunter Tier 1	0.5657	PD	14	DiegoFusaro	99	bit.ly	7
Dormant Followers	16.0%	bot-hunter Tier 2	0.7706	Renzi	11	IAmJamesTheBond	85	imolaoggi.it	7
Bot Followers	24.6%	bot-hunter Tier 3	0.686	Aquarius	10	CesareSacchetti	60	liberoquotidiano.it	6
Avg Mentions	1.379	botometer	0.1481	PonteMorandi	10	Vickiegogreen	54	voxnews.info	6
Retweets RU Prop.	True	Debot	False	pd	9	GiancarloDeRisi	43	corriere.it	6

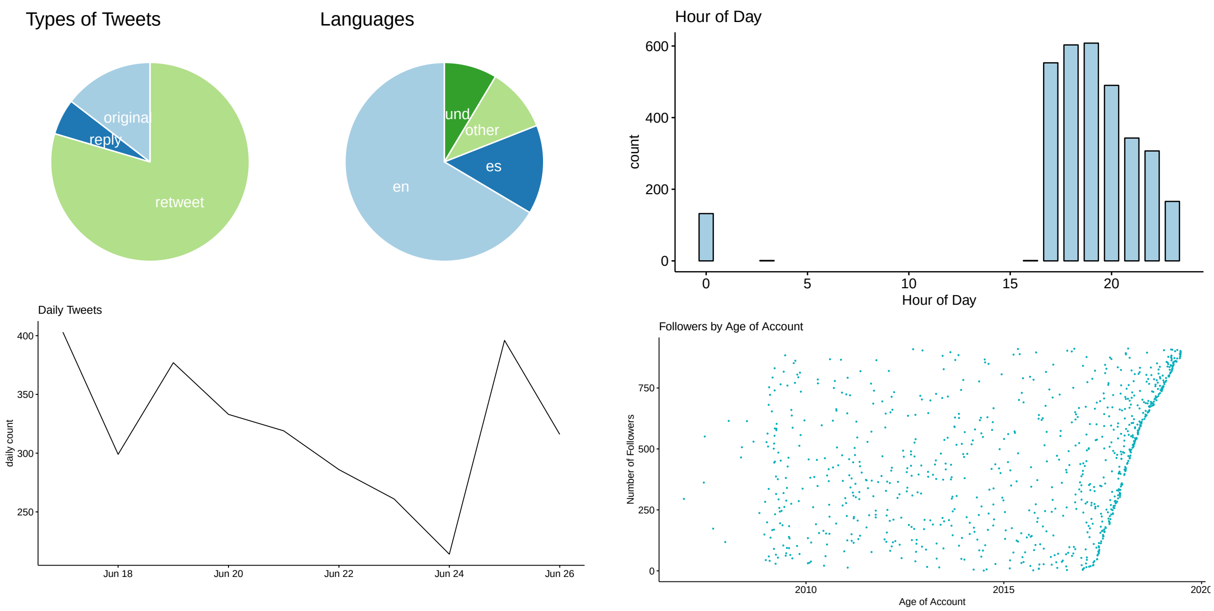


„I don't know what your convictions are worth if you aren't ready to suffer for them or even to die."

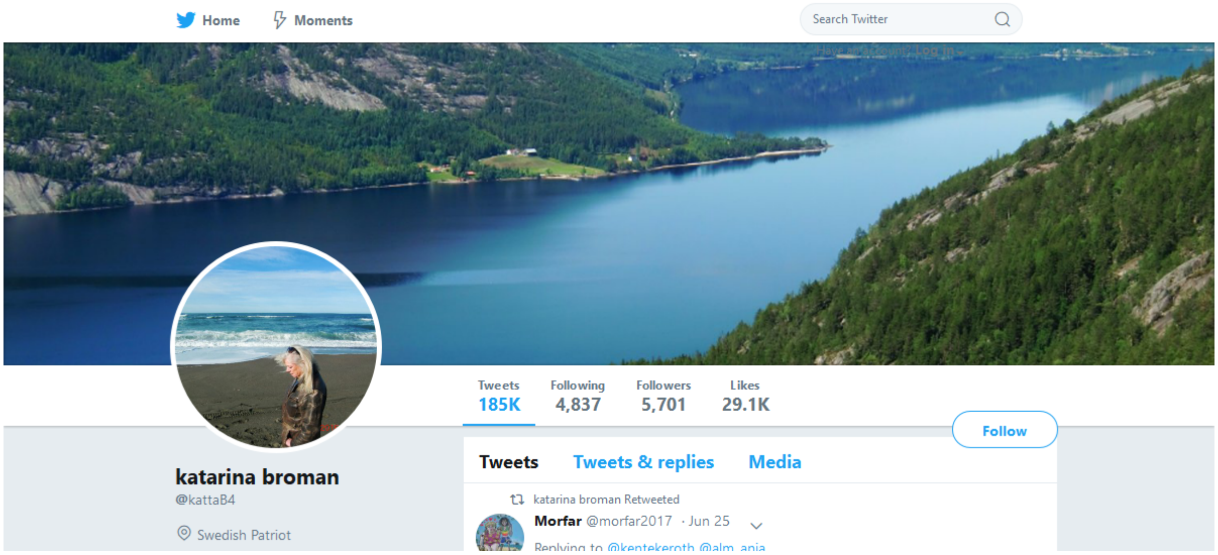
jeff franzius
@jk55044
Joined June 2014

jeff franzius Retweeted
House Foreign Affairs GOP @HouseForeignGOP · 21h
Our commitment to provide #Taiwan with defensive systems will never be up for

G. @jk55044. We found @jk55044 linked to BALTOPS with an extremely high volume of tweets. This bot is unique in that it appears to amplify voices from America, Russia, and Ukraine. Bots like this are sometimes trying to embed in these respective networks.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	5.481	bot-hunter Tier 0	False	UNSC	41	jguaido	122	bit.ly	37
Avg Sentiment (0-1)	0.658	bot-hunter Tier 1	0.7578	NATO	36	SecPompeo	56	pscp.tv	15
Dormant Followers	24.9%	bot-hunter Tier 2	0.6703	Venezuela	35	AsambleaVE	53	youtu.be	11
Bot Followers	28.7%	bot-hunter Tier 3	0.614	Russia	32	mbachelet	51	europa.eu	8
Avg Mentions	1.436	botometer	0.2838	Syria	28	UKUN_NewYork	51	ow.ly	8
Retweets RU Prop.	True	Debot	False	SOHR	26	franceonu	46	hill.cm	7


Home Moments

Search Twitter

Tweets 185K Following 4,837 Followers 5,701 Likes 29.1K

Follow

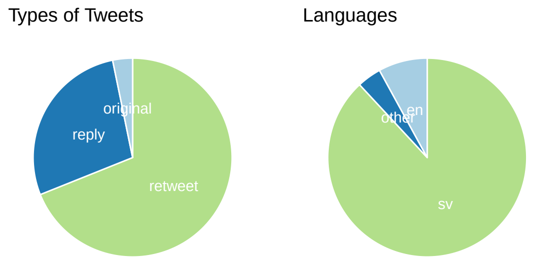
katarina broman
@kattaB4

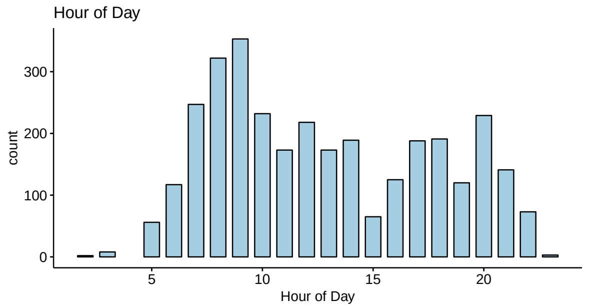
Swedish Patriot

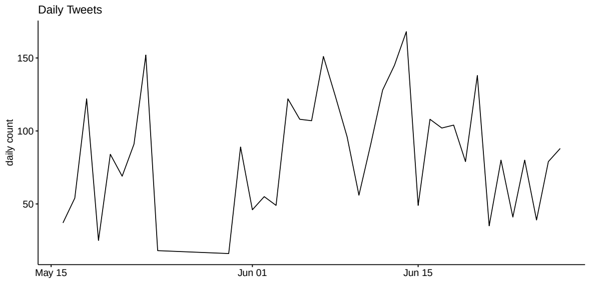
Tweets Tweets & replies Media

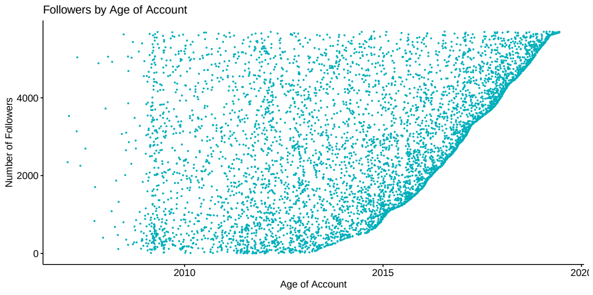
katarina broman Retweeted
Morfar @morfar2017 · Jun 25
Replying to @kentekeroth @alm_ania

H. @kattaB4. We found @kattaB4 as a prolific bot in the Swedish Elections. This account is amplifying some of the cyborg accounts.

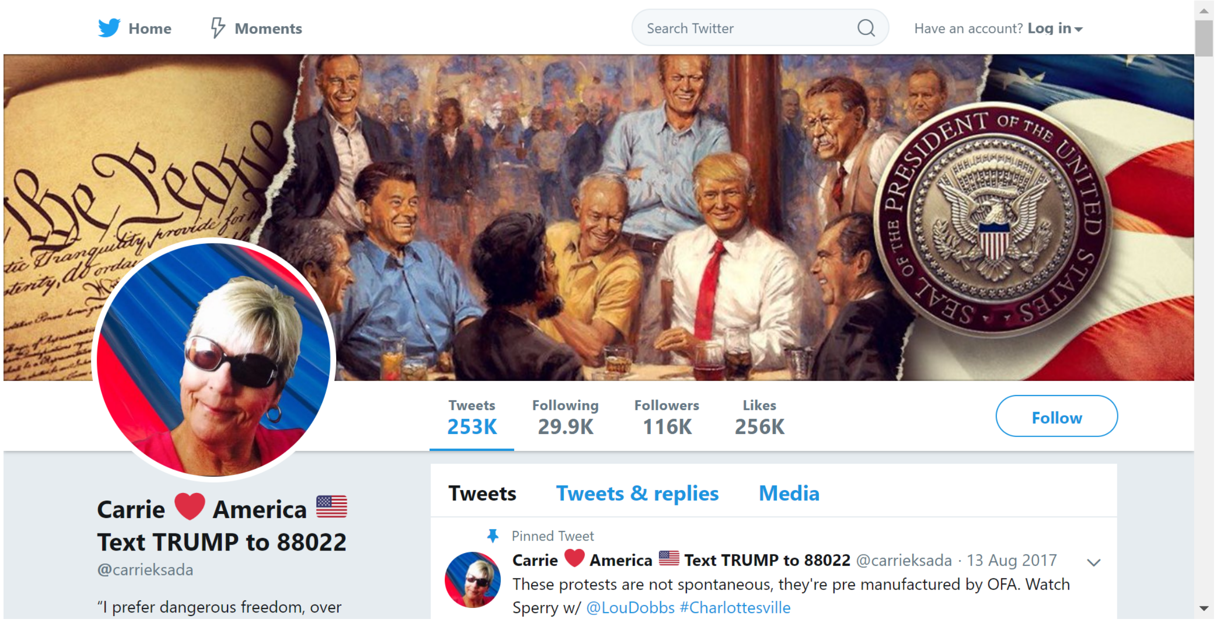








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.848	bot-hunter Tier 0	False	svpol	131	kattaB4	187	svt.se	26
Avg Sentiment (0-1)	0.59	bot-hunter Tier 1	0.6497	migpol	19	BengtHojer	118	samnytt.se	25
Dormant Followers	32.5%	bot-hunter Tier 2	0.8694	Trump	7	TommyFunebo	101	expressen.se	17
Bot Followers	36.7%	bot-hunter Tier 3	0.542	Expressen	7	Ericson_ubbhult	85	aftonbladet.se	14
Avg Mentions	1.612	botometer	0.1481	SVT	4	katjanouch	67	nyheteridag.se	8
Retweets RU Prop.	True	Debot	False	TV4	4	svtnyheter	61	youtu.be	8



Home Moments Search Twitter Have an account? Log in

Tweets 253K Following 29.9K Followers 116K Likes 256K Follow

Carrie ❤ America 🇺🇸 Text TRUMP to 88022
@carrieksada

"I prefer dangerous freedom, over

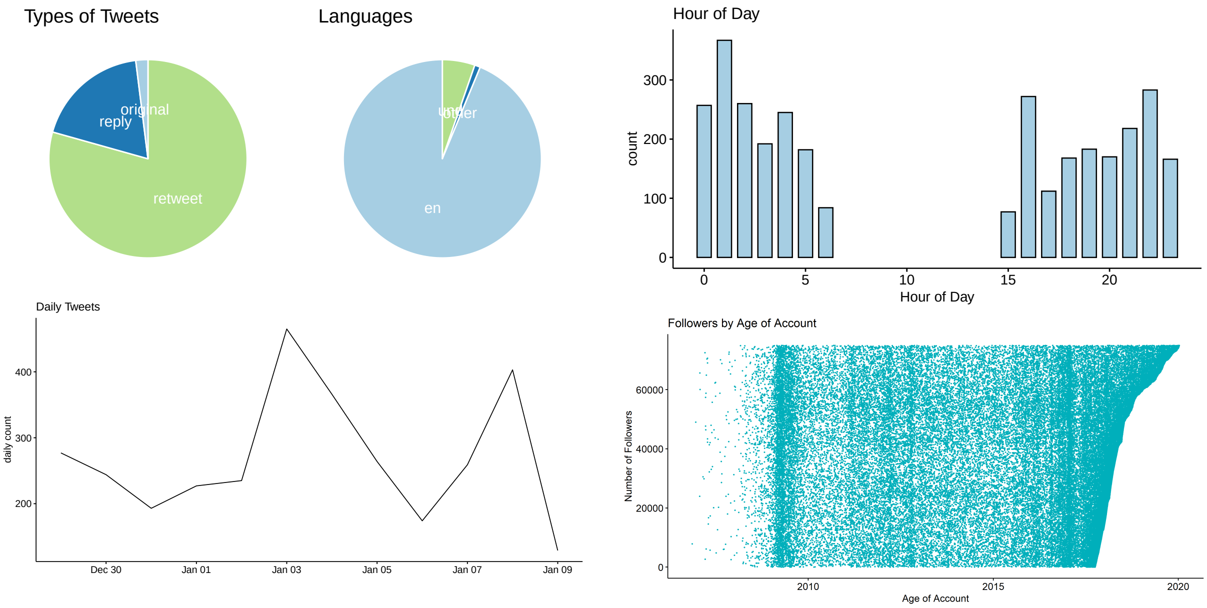
Tweets Tweets & replies Media

Pinned Tweet

Carrie ❤ America 🇺🇸 Text TRUMP to 88022 @carrieksada · 13 Aug 2017

These protests are not spontaneous, they're pre manufactured by OFA. Watch Sperry w/ @LouDobbs #Charlottesville

I. @carrieksada. We found @carrieksada is a popular bot that amplifies the political right in the USA.

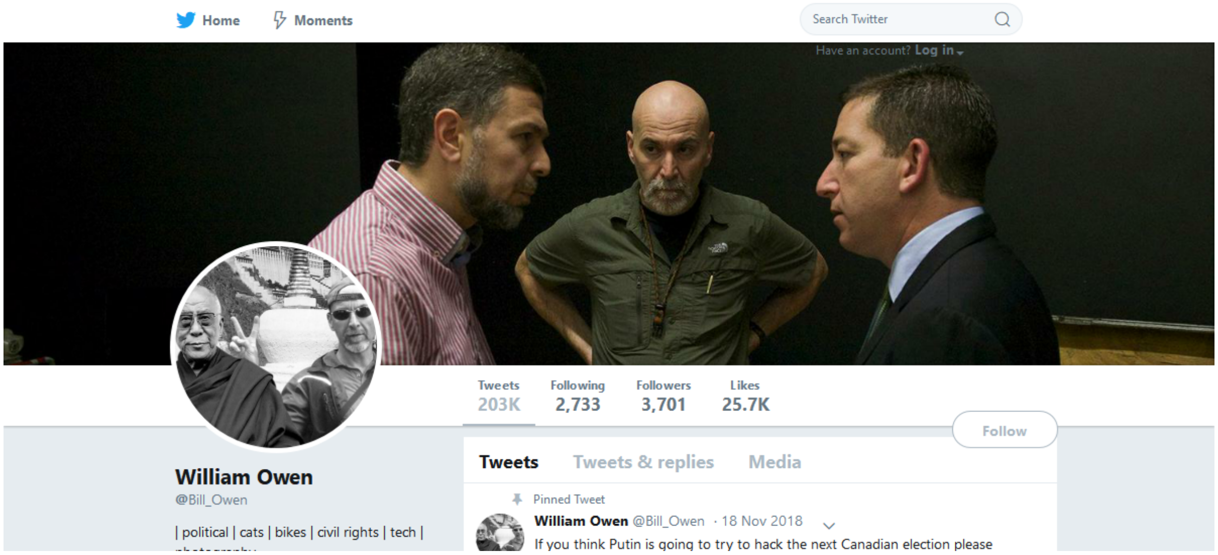


Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.259	bot-hunter Tier 0	False	*Iran*	27	realDonaldTrump	596	dailywire.com	11
Avg Sentiment (0-1)		bot-hunter Tier 1	0.5241	*Soleimani*	21	carrieksada	366	juliereichwein.info	7
Dormant Followers	26.6%	bot-hunter Tier 2	0.8844	*ITSELF*	14	michaelbeatty3	110	foxnews.com	5
Bot Followers	28.2%	bot-hunter Tier 3	0.409	*AND*	11	steph93065	78	buff.ly	5
Avg Mentions	1.982	botometer	0.4745	*TehranNancy*	9	ROHLL5	71	bit.ly	4
Retweets RU Prop.	False	Debot	False	*BREAKING*	9	POTUS	61	bongino.com	4

E.3 Cyborg and Troll accounts

While we often try to force a binary classification of *bot* or *human*, in reality the same Twitter account can have both human activity and computer activity. Accounts that exhibit both human and computer activity are often called *cyborg* or *hybrid* accounts. On the far end of the spectrum of human activity, when a malicious or propaganda account has all human activity, it is often called a *troll* account.

Human activity on an account is often identified in nuanced dialogue, often in reply threads. Computer activity on these same accounts is often identified by the scale and timing of other messages. The computer often conducts amplification tasks at scale, while the human conducts the nuanced dialogue.



Tweets 203K
Following 2,733
Followers 3,701
Likes 25.7K

William Owen
@Bill_Owen

| political | cats | bikes | civil rights | tech |

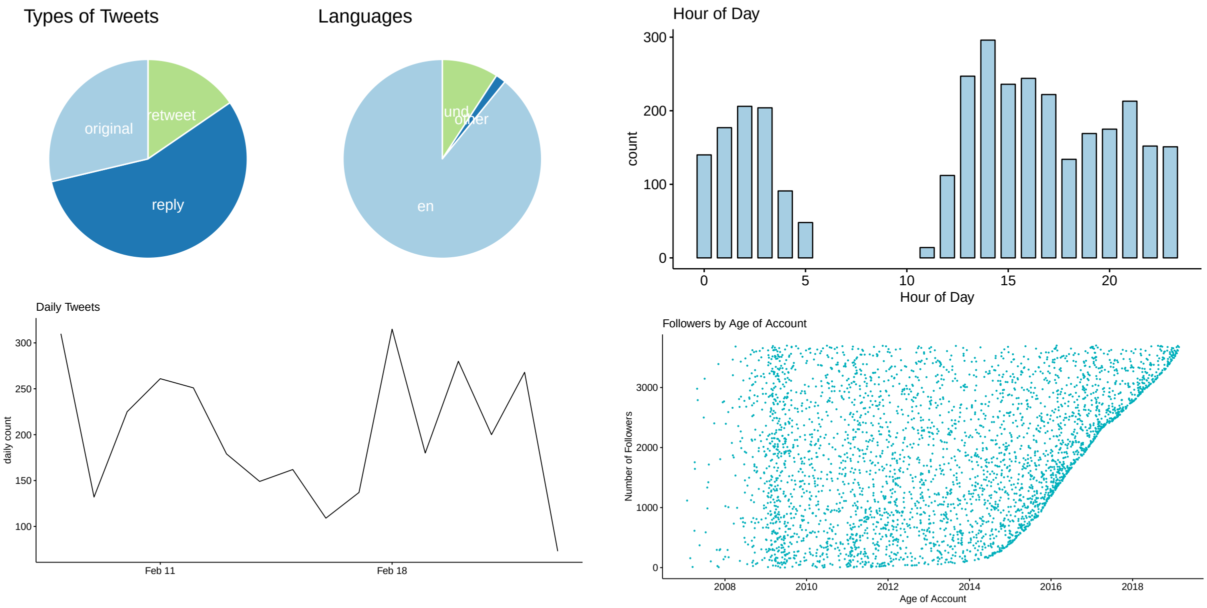
Tweets | Tweets & replies | Media

Pinned Tweet

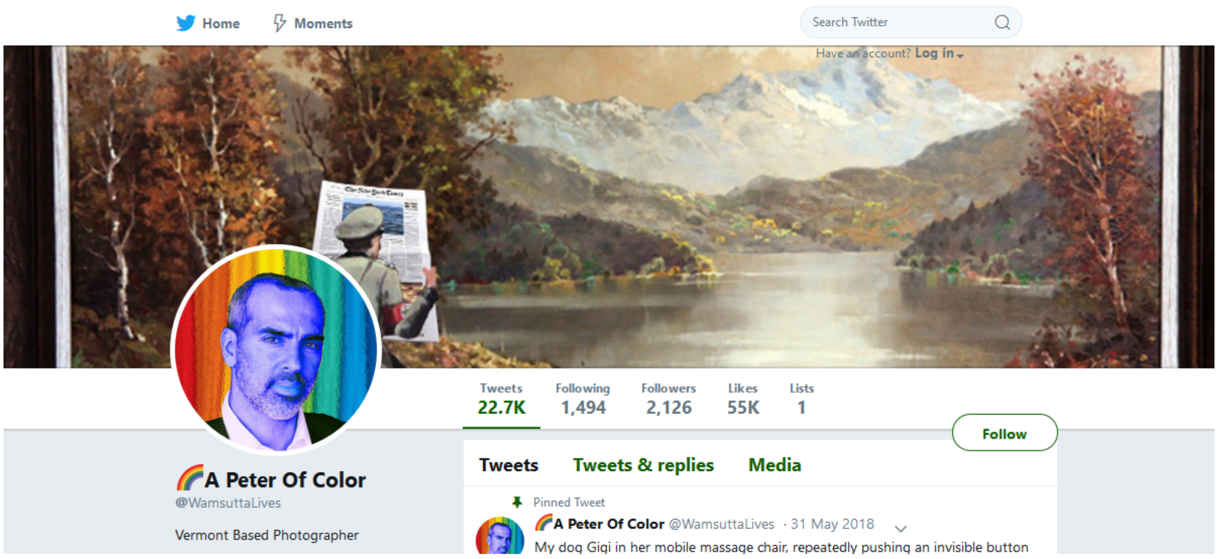
William Owen @Bill_Owen · 18 Nov 2018

If you think Putin is going to try to hack the next Canadian election please

J. @Bill_Owen. @Bill_Owen was discovered in the Trident Juncture data and appears to be a sophisticated *cyborg* account with possible links to Russia. Content focused on US politics and international events. High volume with nuanced dialogue.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.735	bot-hunter Tier 0	False	PresidentsDay	16	ggreenwald	292	youtu.be	89
Avg Sentiment (0-1)	0.57	bot-hunter Tier 1	0.7642	Hotthawa	8	PresumptuousBug	206	wikipedia.org	6
Dormant Followers	27.2%	bot-hunter Tier 2	0.824	Venezuela	8	trapdinawrpool	99	markdanner.com	6
Bot Followers	45.5%	bot-hunter Tier 3	0.555	FunFacts	7	Bill_Owen	88	cbc.ca	6
Avg Mentions	1.806	botometer	0.0765	yearzeroamerica	7	shenebraskan	81	interc.pt	5
Retweets RU Prop.	True	Debot	False	PTUIN	7	Pedinska	81	fair.org	5



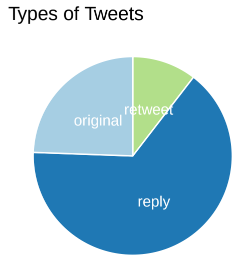
Tweets	Following	Followers	Likes	Lists
22.7K	1,494	2,126	55K	1

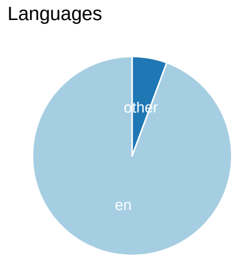
🌈A Peter Of Color
@WamsuttaLives

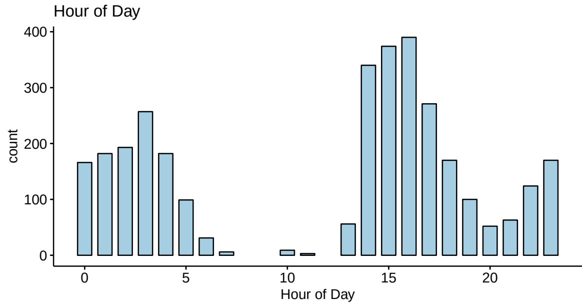
Vermont Based Photographer

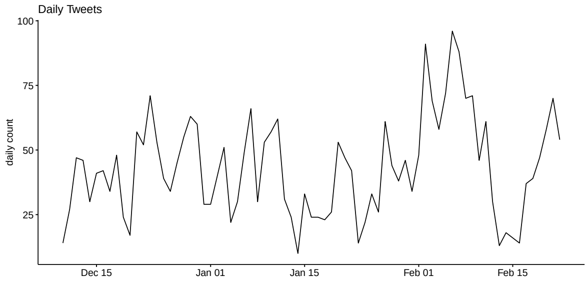
Pinned Tweet
🌈A Peter Of Color @WamsuttaLives · 31 May 2018
My dog Gigi in her mobile massage chair, repeatedly pushing an invisible button

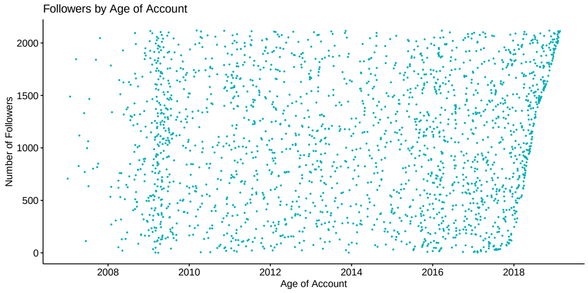
K. @WamsuttaLives. We found @WamsuttaLives linked to @Bill_Owen. Note the @Bill_Owen appears to be a sophisticated *cyborg* account with highly possible links to Russia. @WamsuttaLives also appears to be a sophisticated cyborg or full time troll. Content focused on US politics and international events. High volume with nuanced dialogue.



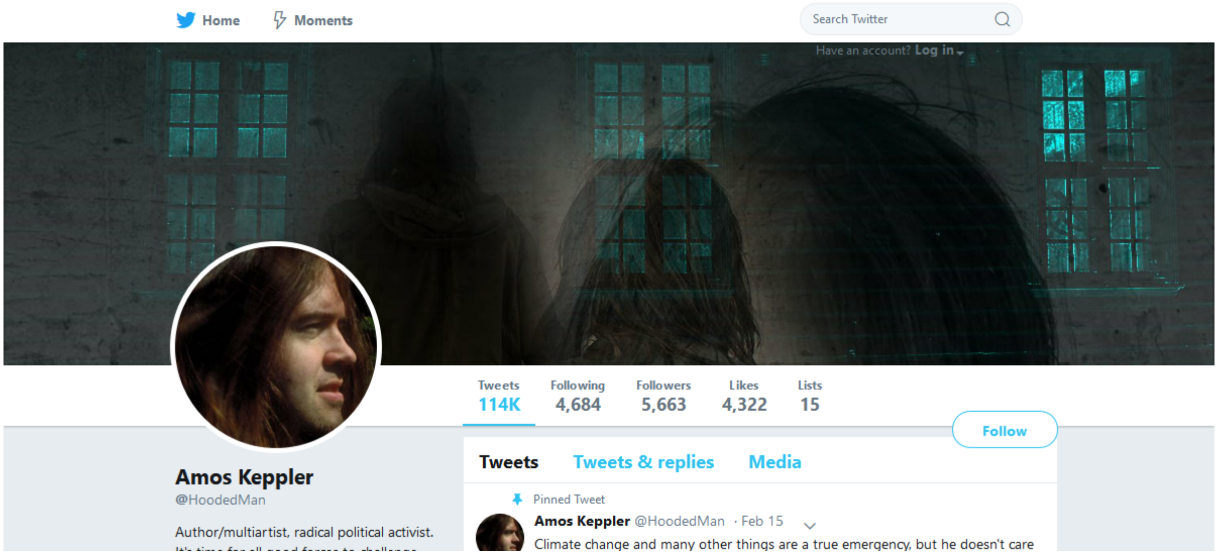








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.742	bot-hunter Tier 0	False	YellowVests	3	ggreenwald	101	youtube.com	35
Avg Sentiment (0-1)	0.571	bot-hunter Tier 1	0.8782	YankeeGoHome	2	aaronjmate	81	youtu.be	9
Dormant Followers	5.7%	bot-hunter Tier 2	0.6679	IntegrityInitiative	2	BethLynch2020	72	huffingtonpost.com	7
Bot Followers	59.6%	bot-hunter Tier 3	0.621	Russian	2	Bill_Owen	62	theintercept.com	7
Avg Mentions	1.96	botometer	0.1285	Venezuela	2	AOC	44	patreon.com	4
Retweets RU Prop.	False	Debot	False	MoveLeftIdiots	2	bourgeoisalien	42	dailykos.com	4



Home Moments Search Twitter

Have an account? Log in

Tweets 114K Following 4,684 Followers 5,663 Likes 4,322 Lists 15

Follow

Amos Keppler
@HoodedMan

Author/multiartist, radical political activist.

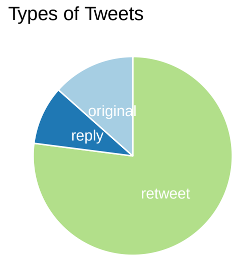
Tweets Tweets & replies Media

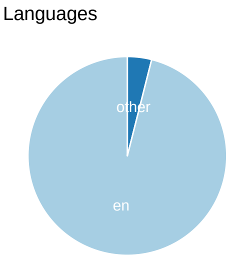
Pinned Tweet

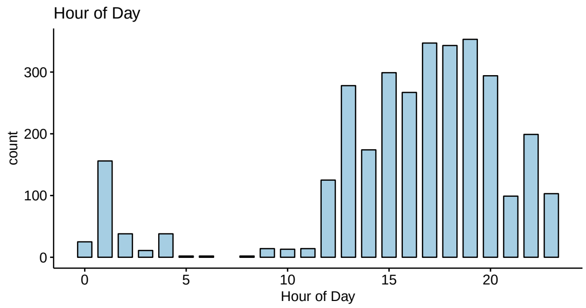
Amos Keppler @HoodedMan · Feb 15

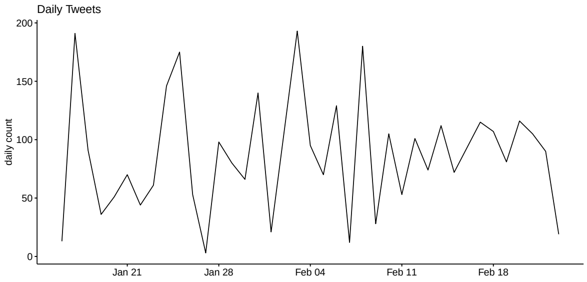
Climate change and many other things are a true emergency, but he doesn't care

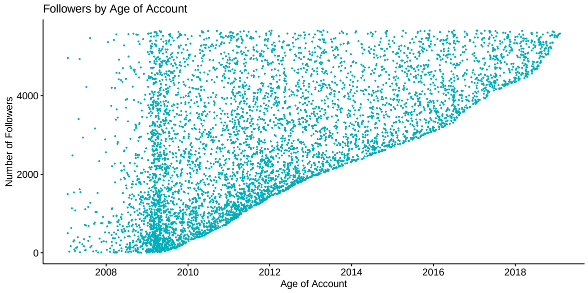
L. @HoodedMan. HoodedMan was discovered in the Trident Juncture data, and was an adamant opponent of Trident Juncture, to include posting pictures appearing to organize protests against it. This account has extremely high volume and high retweets, with a decent likelihood of having some level of automation. This is most likely a cyborg account.



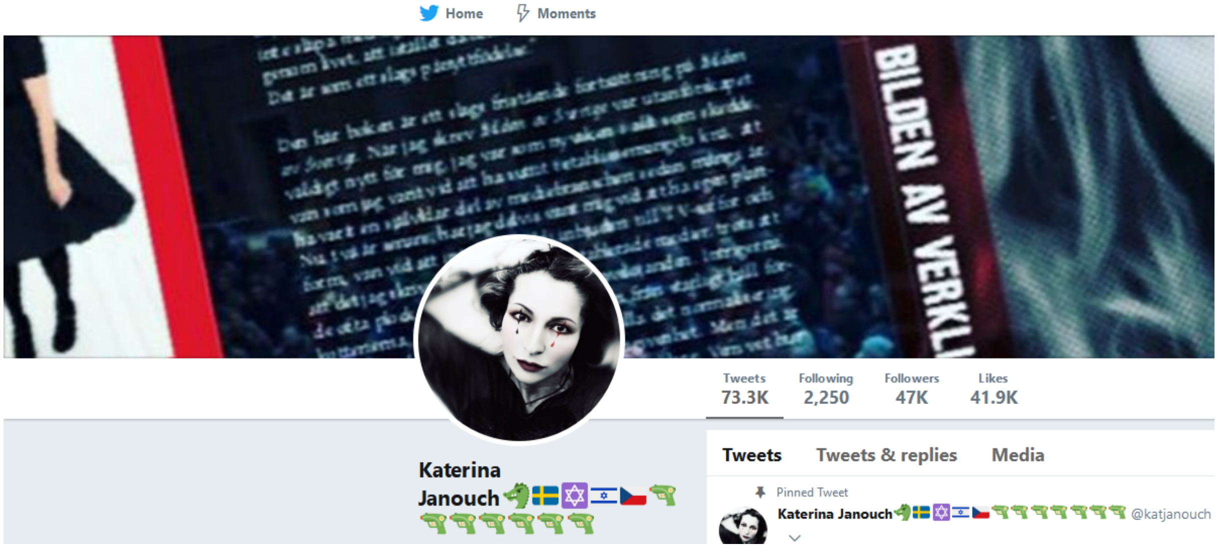








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.84	bot-hunter Tier 0	False	Venezuela	24	HoodedMan	147	blogspot.com	134
Avg Sentiment (0-1)	0.572	bot-hunter Tier 1	0.6423	amwriting	15	AdriMoreau	127	theguardian.com	19
Dormant Followers	25.4%	bot-hunter Tier 2	0.6972	bbcqt	13	AsterdisPrime	121	midnight-fire.net	14
Bot Followers	39.7%	bot-hunter Tier 3	0.613	Gaza	10	BlueWatchman	89	skwawkbox.org	13
Avg Mentions	1.755	botometer	0.2363	US	9	TheWrongWoman	89	independent.co.uk	11
Retweets RU Prop.	True	Debot	False	Israeli	8	ColeslamTalia	57	bit.ly	11



Home Moments

BILDEN AV VERKLI

Tweets	Following	Followers	Likes
73.3K	2,250	47K	41.9K

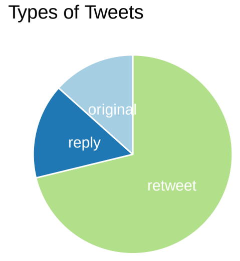
Katerina Janouch🐊🇸🇪🔯🇮🇱🇨🇿🔫🔫🔫🔫🔫🔫🔫

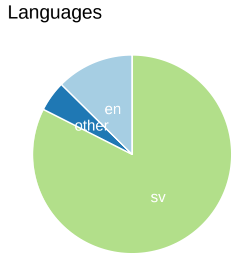
Tweets Tweets & replies Media

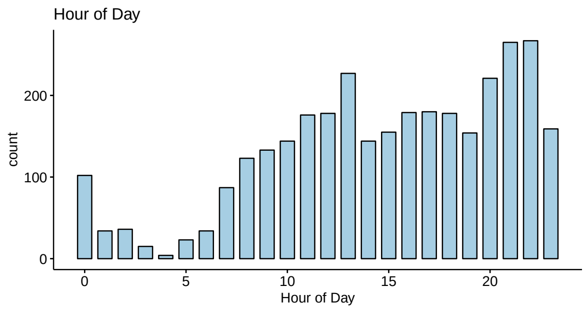
📌 Pinned Tweet

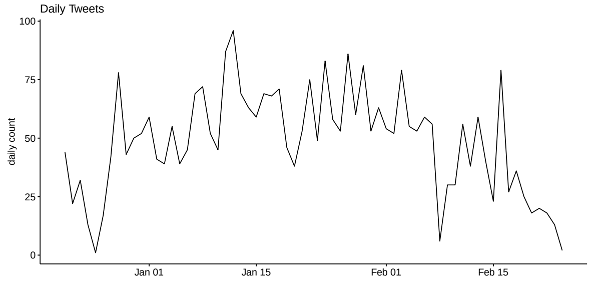
Katerina Janouch🐊🇸🇪🔯🇮🇱🇨🇿🔫🔫🔫🔫🔫🔫🔫 @katjanouch

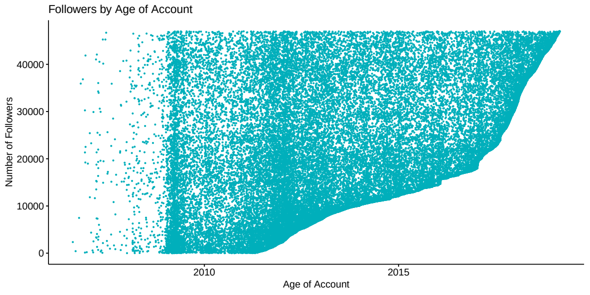
M. @katjanouch. @katjanouch is the top mentioned agent and #1 most influential account in the Swedish 2018 election. The account belongs to a Czech born Swedish Journalist who was criticized for spreading Russian Propaganda. Her Twitter account recently averages 62 Tweets per day. Her account supports Matteo Salvini (an Italian politician who is a Euro sceptic), Ari Fuld (an American Israeli settler with alleged connections to Alt-right movements), and the far right website "Voice of Europe"



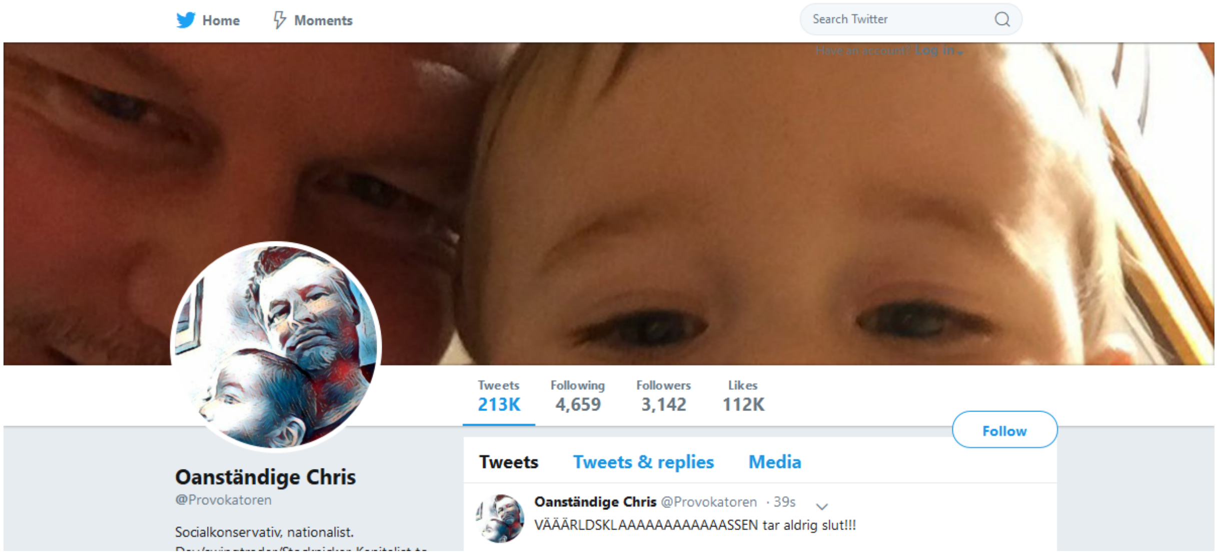








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.047	bot-hunter Tier 0	False	svpol	254	katjanouch	183	samnytt.se	45
Avg Sentiment (0-1)	0.587	bot-hunter Tier 1	0.796	migpol	76	sjunnedotcom	97	nyheteridag.se	32
Dormant Followers	54.1%	bot-hunter Tier 2	0.8428	GiletsJaunes	31	MickeK69	82	expressen.se	31
Bot Followers	58.8%	bot-hunter Tier 3	0.353	projektsanning	29	thereseverdun	66	svt.se	25
Avg Mentions	1.405	botometer	0.0608	France	13	V_of_Europe	65	katerinamagasin.se	23
Retweets RU Prop.	False	Debot	False	Macron	10	PeterSellei	62	aftonbladet.se	20



Tweets	Following	Followers	Likes
213K	4,659	3,142	112K

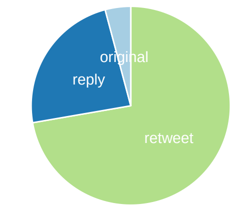
Oanständige Chris
@Provokatoren

Socialkonservativ, nationalist.

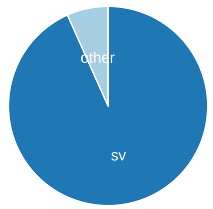
Oanständige Chris @Provokatoren · 39s
VÄÄÄRLDSKLAAAAAAAAAAAAASSEN tar aldrig slut!!!

N. @Provokatoren. @Provokatoren is a bot that became #4 in Betweenness Score for Swedish Elections This bot's goal is to bridge groups. It recently averaged 175 Tweets per day. Many tweets are anti-immigrant.

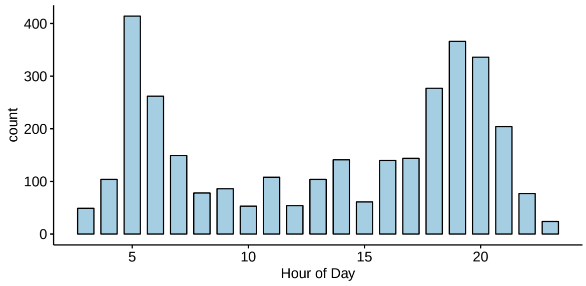
Types of Tweets



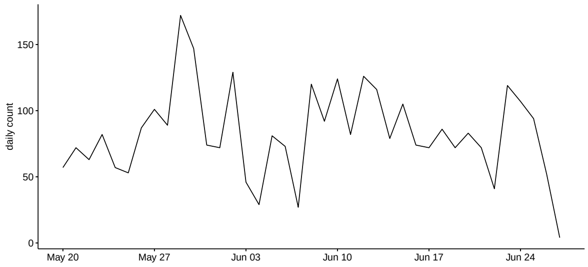
Languages



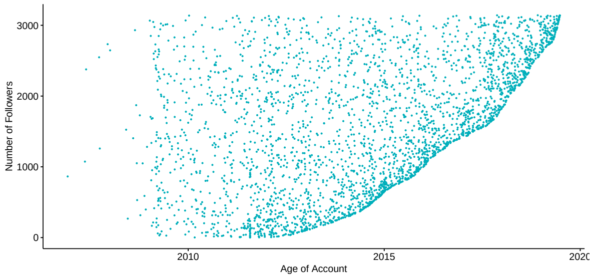
Hour of Day



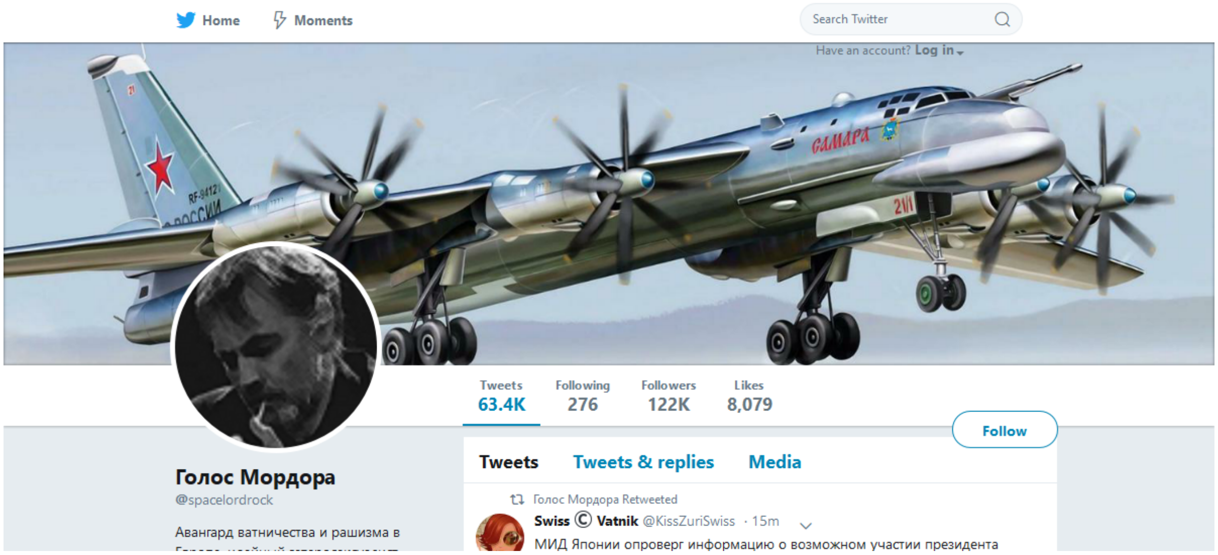
Daily Tweets



Followers by Age of Account



Metric	Value	Model	Probability	hashtag	count	Mention	Count	domain	count
Frd/Fol Ratio	1.483	bot-hunter Tier 0	False	svpol	747	dumskallar	330	samnytt.se	69
Avg Sentiment (0-1)	0.591	bot-hunter Tier 1	0.6528	migpol	119	MickeK69	219	expressen.se	42
Dormant Followers	26.3%	bot-hunter Tier 2	0.85	EUval2019	18	BengtHojer	142	svt.se	34
Bot Followers	33.5%	bot-hunter Tier 3	0.428	projektsanning	13	AlexandraHedbo1	126	aftonbladet.se	18
Avg Mentions	1.427	botometer	0.0765	Svpol	11	ROGSAHL	103	nyheteridag.se	13
Retweets RU Prop.	False	Debot	False	Sverigebilden	10	socialdemokrat	77	ledarsidorna.se	13



Home Moments Search Twitter

Have an account? Log in

Tweets	Following	Followers	Likes
63.4K	276	122K	8,079

Follow

Голос Мордора
@spacelordrock

Авангард ватничества и рашизма в

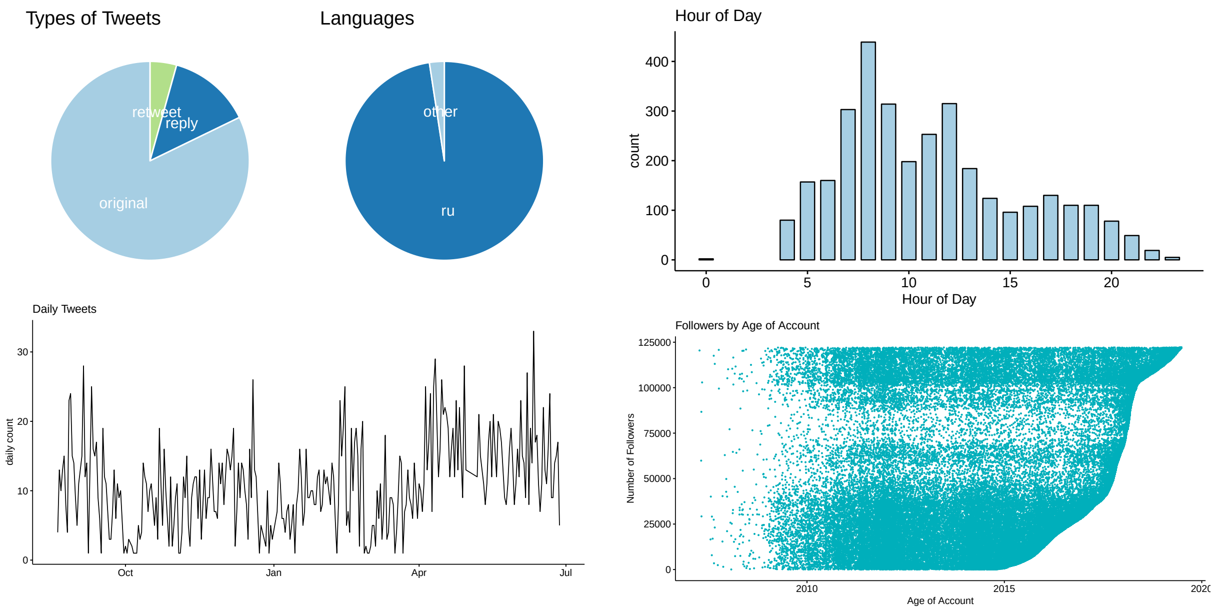
Tweets | Tweets & replies | Media

Голос Мордора Retweeted

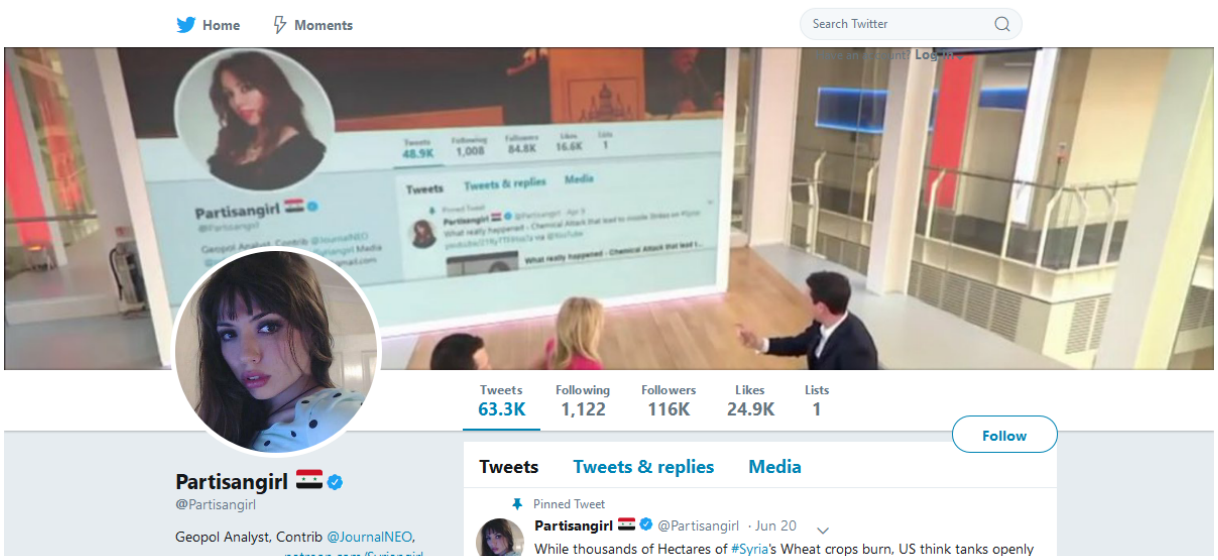
Swiss © Vatnik @KissZuriSwiss · 15m

МИД Японии опроверг информацию о возможном участии президента

O. @spacelordrock. This is a very popular and influential Russian Troll and Propaganda account.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.002	bot-hunter Tier 0	False	*Victory Day*	2	KissZuriSwiss	16	youtu.be	7
Avg Sentiment (0-1)	0.638	bot-hunter Tier 1	0.1558	*JulianAssange*	1	rlz$_the_kraken$	13	ria.ru	6
Dormant Followers	61.0%	bot-hunter Tier 2	0.4834	*Donetsk*	1	Vityzeva	11	youtube.com	4
Bot Followers	61.0%	bot-hunter Tier 3	0.335	*WWII*	1	friendlybus	11	google.nl	2
Avg Mentions	1.112	botometer	0.052	*9maya*	1	roughbear	10	news-front.info	2
Retweets Russia Propaganda	True	Debot	False	*FreeAssange*	1	spacelordrock	9	amedia.press	2



Tweets 63.3K | Following 1,122 | Followers 116K | Likes 24.9K | Lists 1

Follow

Partisangirl

@Partisangirl

Geopol Analyst, Contrib @JournalNEO,

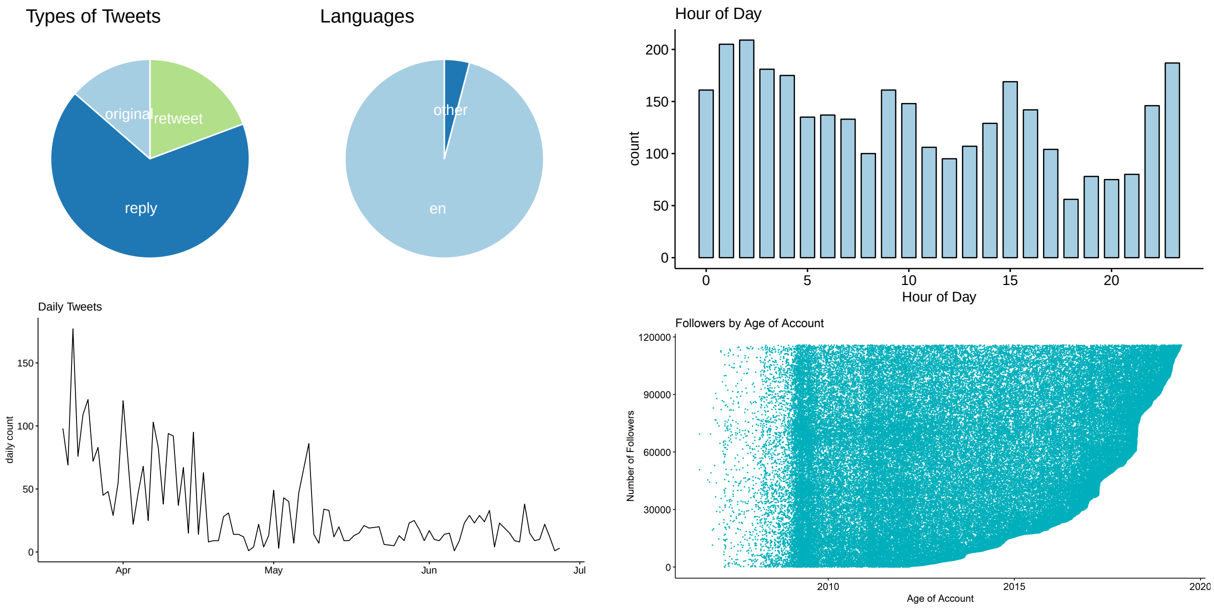
Tweets | Tweets & replies | Media

Pinned Tweet

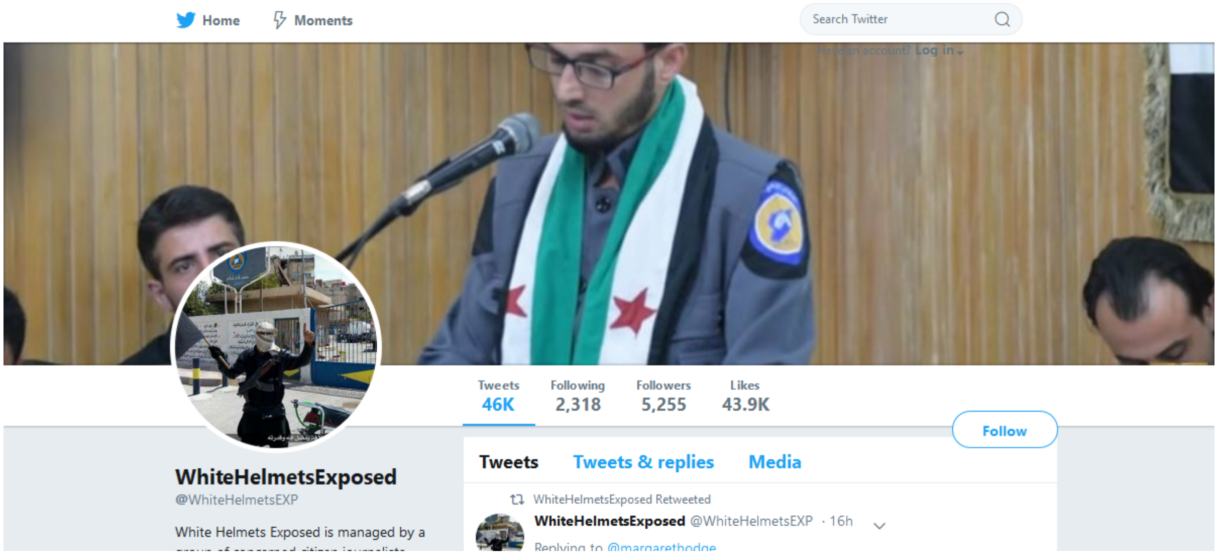
Partisangirl @Partisangirl · Jun 20

While thousands of Hectares of #Syria's Wheat crops burn, US think tanks openly

P. @Partisangirl. We found @Partisangirl in numerous articles. This is a well known Russian/Syrian troll account that the Daily Beast called “The Kardashian Look-Alike Trolling for Assad”.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.01	bot-hunter Tier 0	False	Syria	155	Partisangirl	233	youtu.be	37
Avg Sentiment (0-1)	0.55	bot-hunter Tier 1	0.7592	Israel	65	21WIRE	63	youtube.com	19
Dormant Followers	23.7%	bot-hunter Tier 2	0.7259	GolanHeights	60	PrisonPlanet	52	google.com.au	10
Bot Followers	39.4%	bot-hunter Tier 3	0.491	Idlib	36	CassandraRules	49	pscp.tv	6
Avg Mentions	2.166	botometer	0.0657	Trump	27	KevorkAlmassian	49	wikipedia.org	6
Retweets RU Prop.	True	Debot	False	Assange	26	GregoryPWaters	47	rt.com	5



WhiteHelmetsExposed
@WhiteHelmetsEXP

White Helmets Exposed is managed by a

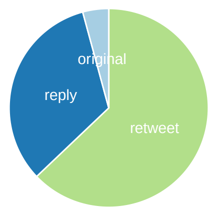
WhiteHelmetsExposed Retweeted

WhiteHelmetsExposed @WhiteHelmetsEXP · 16h

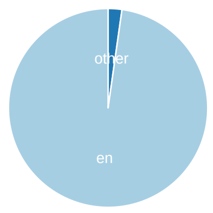
Replying to @margarethodge

Q. @WhiteHelmetsEXP. The @WhiteHelmetsEXP cyborg account attacks the so-called white helmets. According to Wikipedia, "The White Helmets, officially known as Syria Civil Defence, is a volunteer organisation that operates in parts of rebel-controlled Syria and in Turkey."

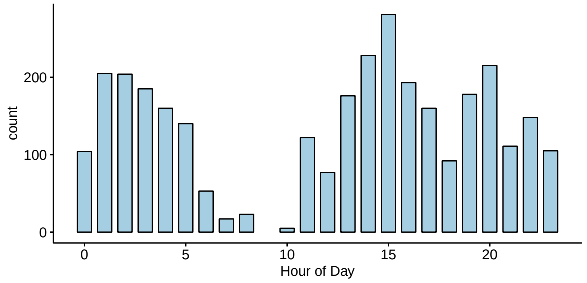
Types of Tweets



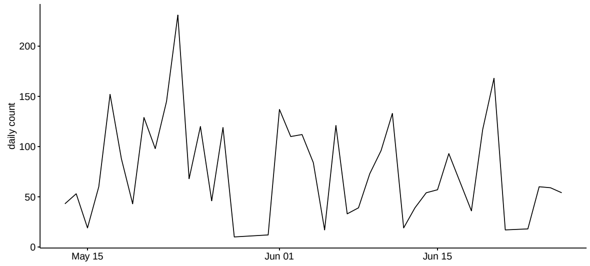
Languages



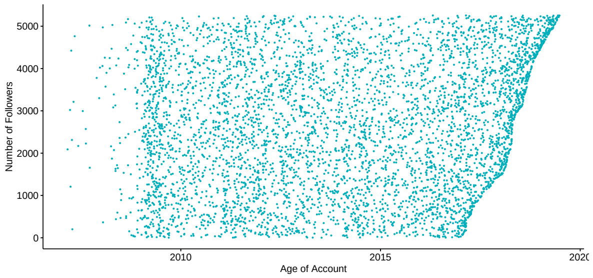
Hour of Day



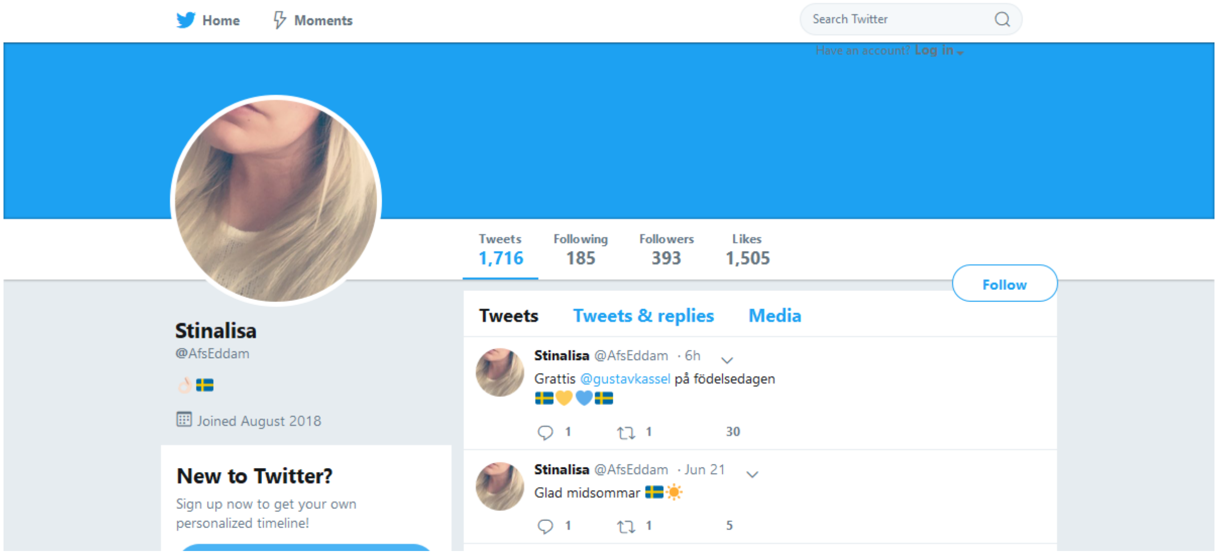
Daily Tweets



Followers by Age of Account



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.441	bot-hunter Tier 0	False	AlQaeda	582	WhiteHelmetsEXP	272	youtube.com	31
Avg Sentiment (0-1)	0.601	bot-hunter Tier 1	0.6765	WhiteHelmets	417	realDonaldTrump	136	rt.com	12
Dormant Followers	10.9%	bot-hunter Tier 2	0.7627	Israel	148	GeorgeMonbiot	118	mintpressnews.com	10
Bot Followers	38.7%	bot-hunter Tier 3	0.562	Syria	68	ClarkeMicah	114	youtu.be	8
Avg Mentions	1.825	botometer	0.1481	USUKQaeda	49	guardian	105	21stcenturywire.com	6
Retweets RU Prop.	True	Debot	False	Terrorism	48	PiersRobinson1	102	consortiumnews.com	6



Tweets 1,716 | Following 185 | Followers 393 | Likes 1,505

Stinalisa
@AfsEddam
👌🇸🇪

Joined August 2018

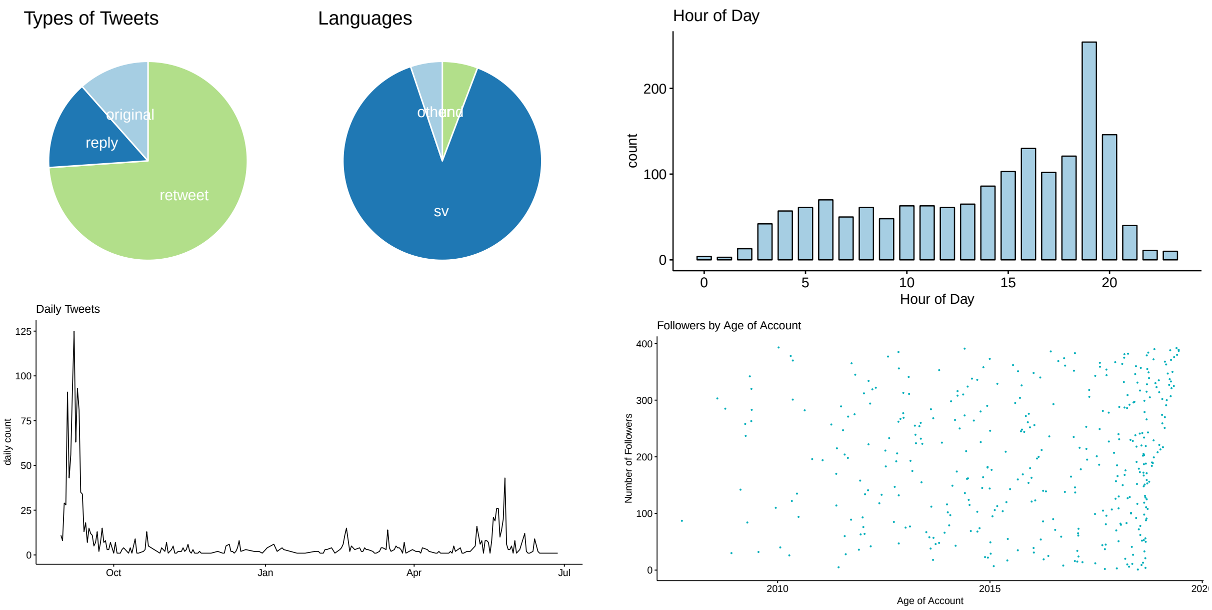
New to Twitter?
Sign up now to get your own personalized timeline!

Tweets | Tweets & replies | Media

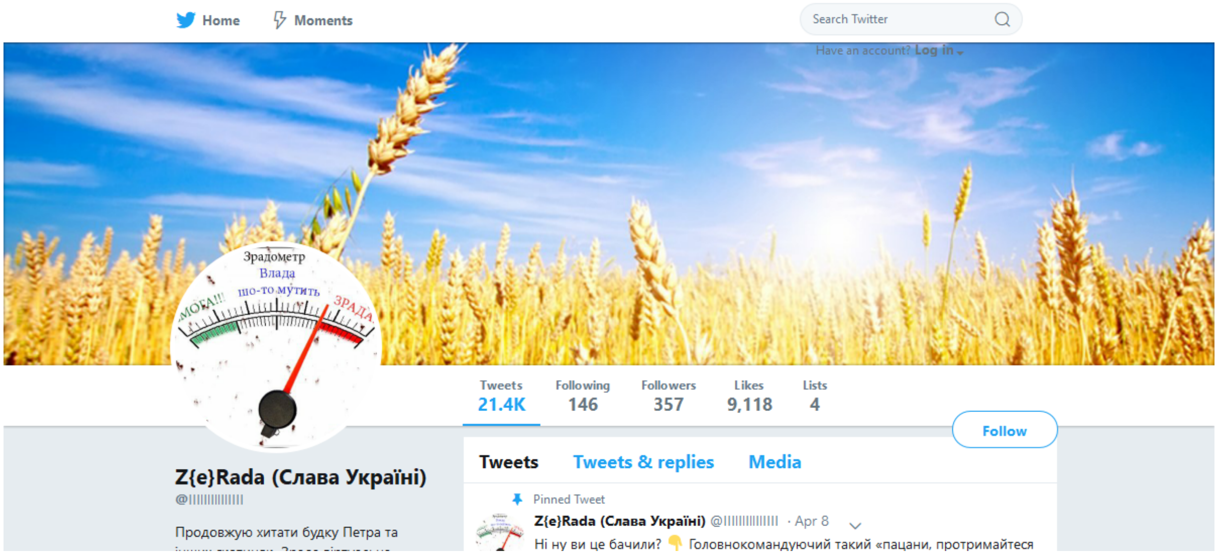
Stinalisa @AfsEddam · 6h
Grattis @gustavkassel på födelsedagen
🇸🇪💛💙🇸🇪
1 1 30

Stinalisa @AfsEddam · Jun 21
Glad midsommar 🇸🇪☀️
1 1 5

R. @AfsEddam. This is a sophisticated cyborg that propped up AFS in Swedish election discussions



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.471	bot-hunter Tier 0	False	svpol	81	AfS_riks	219	youtu.be	27
Avg Sentiment (0-1)	0.609	bot-hunter Tier 1	0.2717	AfS2018	70	gustavkassel	188	samnytt.se	26
Dormant Followers	10.4%	bot-hunter Tier 2	0.4333	AfS2019	40	Jeff_Ahl	79	youtube.com	24
Bot Followers	31.3%	bot-hunter Tier 3	0.303	val2018	24	William_Hahne	56	nyadagbladet.se	13
Avg Mentions	1.318	botometer	0.089	AllaTillKungsan	23	Sambandet	41	pscp.tv	10
Retweets RU Prop.	True	Debot	False	valet2018	15	erikberglund89	40	friatider.se	7



Tweets 21.4K
Following 146
Followers 357
Likes 9,118
Lists 4

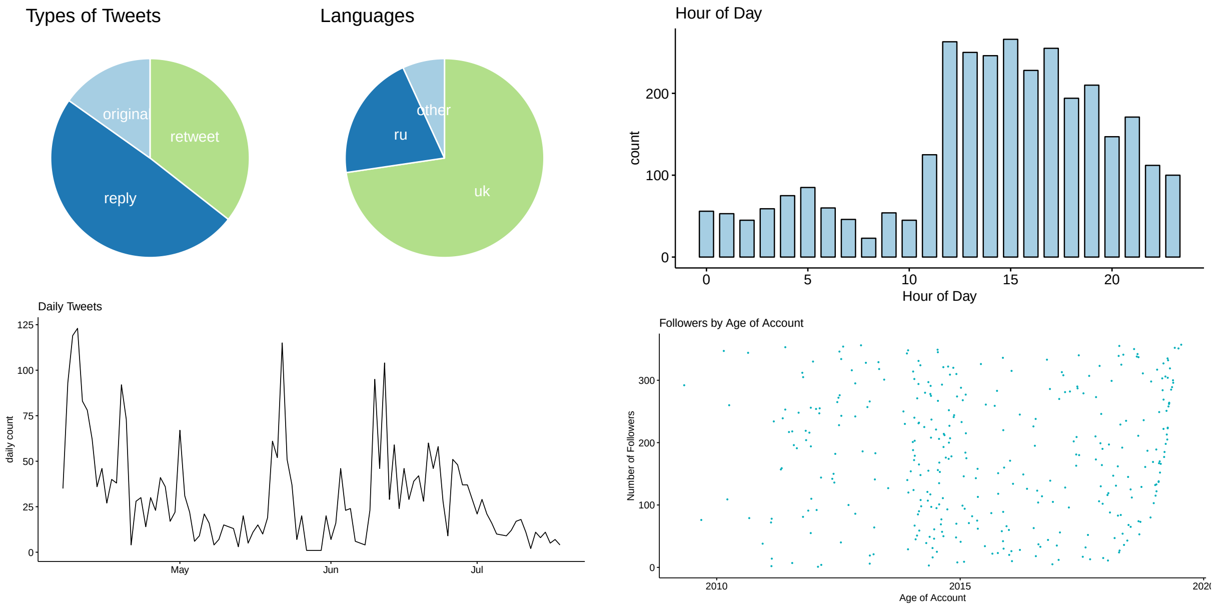
Z{e}Rada (Слава Україні)
@IIIIIIIIIIIII

Продовжую хитати будку Петра та

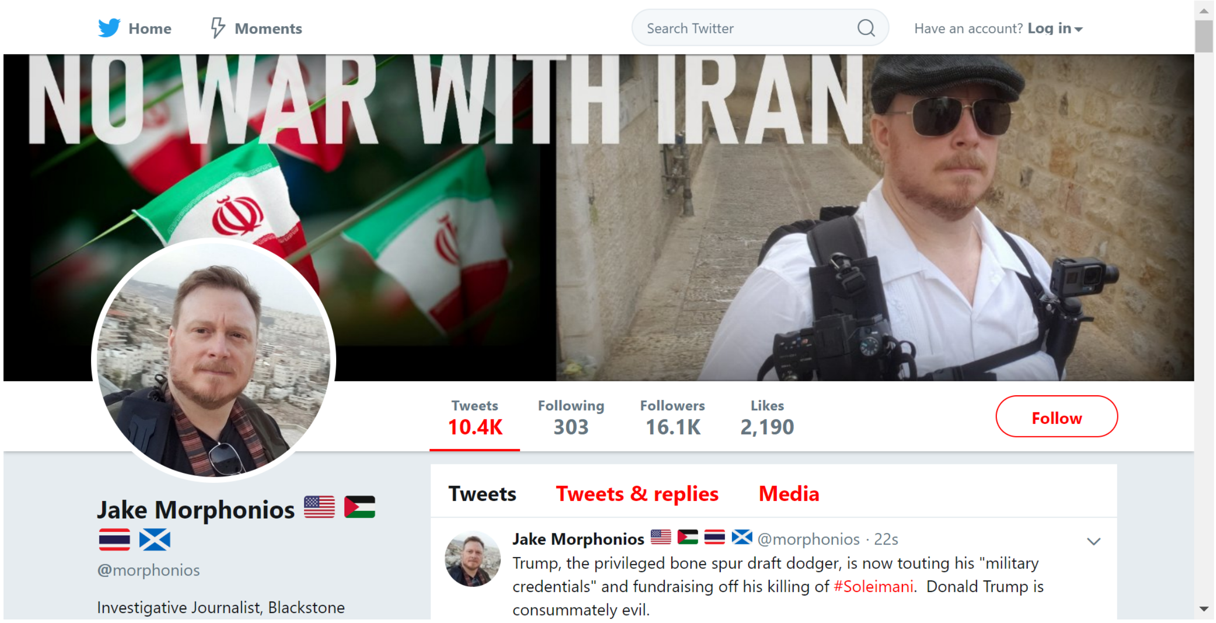
Tweets | Tweets & replies | Media

Pinned Tweet
Z{e}Rada (Слава Україні) @IIIIIIIIIIIII · Apr 8
Ні ну ви це бачили? 👎 Головнокомандуючий такий «пацани, протримайтеся

S. @IIIIIIIIIIIIII. We found @IIIIIIllIIIIII linked to a sophisticated IO campaign against the new President of Ukraine. Of interest is the low score it received from Tier 1 bot-hunter (lowest bot score in the entire conversation). We leveraged bot-match to find bots like it. Location is ‘Santa Clausville, N.P.’



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.409	bot-hunter Tier 0	False	*RuhnovNovovSil*	7	Allehor	220	youtu.be	46
Avg Sentiment (0-1)		bot-hunter Tier 1	0.103	*On the trap*	5	poroshenko	130	facebook.com	5
Dormant Followers	4.8%	bot-hunter Tier 2	0.6195	*Saakashvili*	5	ZelenskyyUa	96	inforesist.org	5
Bot Followers	29.4%	bot-hunter Tier 3	0.335	*shotutdum*	4	zel_prezident	77	youtube.com	4
Avg Mentions	1.599	botometer	0.0709	*APUTINHUILO*	4	NoFated_	60	independent.co.uk	4
Retweets RU Prop.	False	Debot	False	*gunfire*	3	zeteam_official	59	pravda.com.ua	4



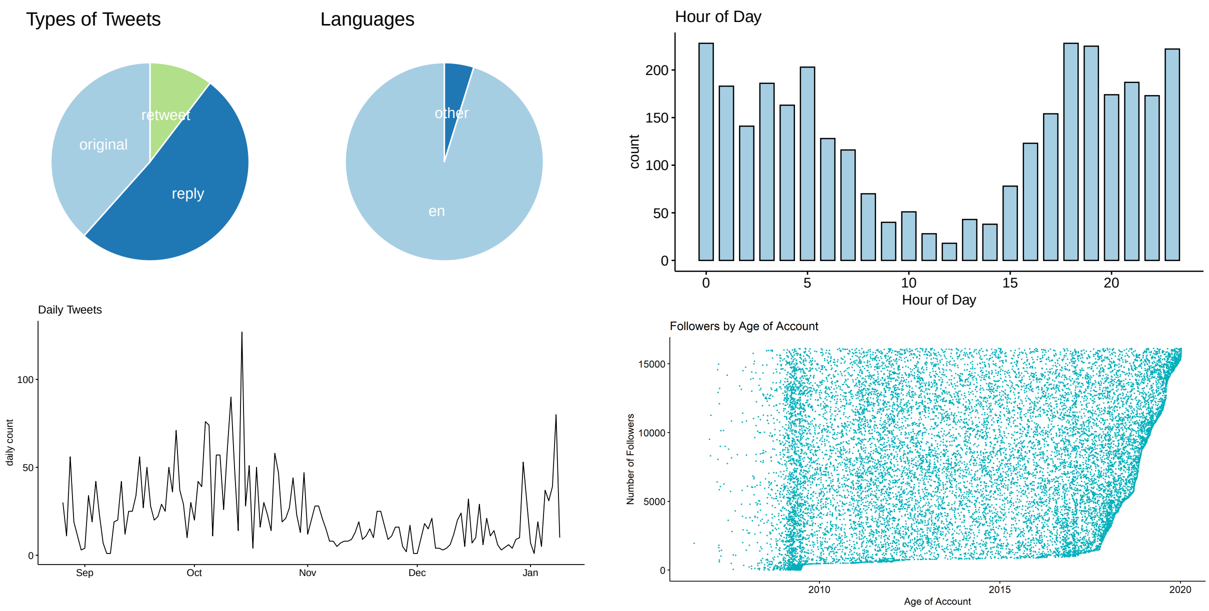
Jake Morphonios 🇺🇸 🇵🇸 🇹🇭 🏴󠁧󠁢󠁳󠁣󠁴󠁿

@morphonios

Investigative Journalist, Blackstone

Jake Morphonios 🇺🇸 🇵🇸 🇹🇭 🏴󠁧󠁢󠁳󠁣󠁴󠁿 @morphonios · 22s
Trump, the privileged bone spur draft dodger, is now touting his "military credentials" and fundraising off his killing of #Soleimani. Donald Trump is consummately evil.

T. @morphonios. We found @morphonios to be a sophisticated cyborg designed to attack American foreign policy.

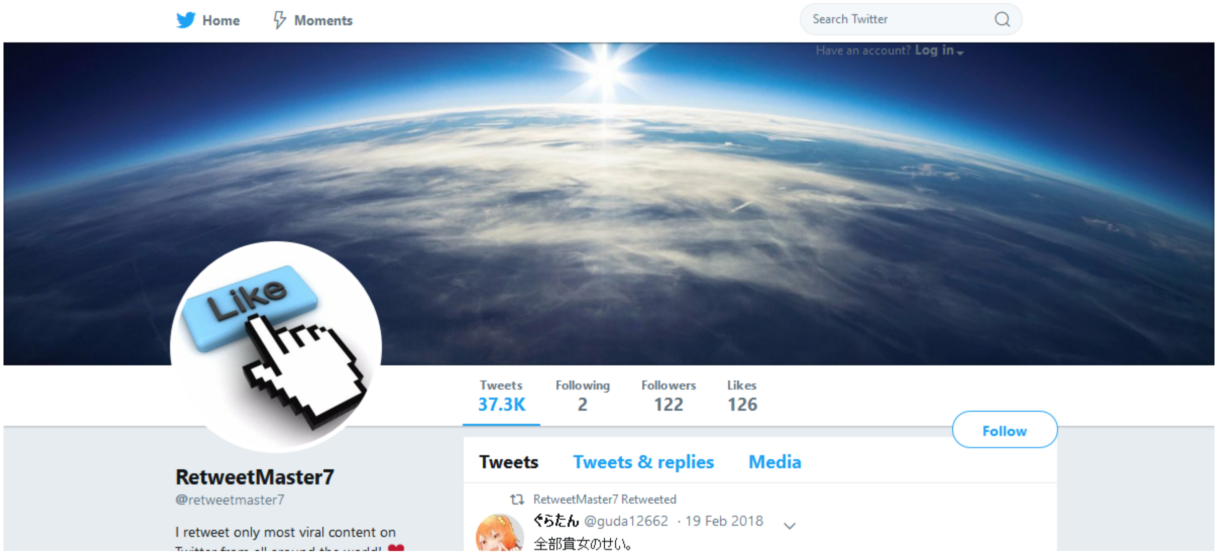


Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.019	bot-hunter Tier 0	False	*Trump*	56	realDonaldTrump	69	youtu.be	32
Avg Sentiment (0-1)		bot-hunter Tier 1	0.4203	*Iran*	45	eRepublicUSA	47	haaretz.com	24
Dormant Followers	21.4%	bot-hunter Tier 2	0.4935	*BDS*	26	OfficialWHPress	40	youtube.com	15
Bot Followers	35.7%	bot-hunter Tier 3	0.343	*Iraq*	17	KathrynRTitus11	28	myshopify.com	8
Avg Mentions	1.225	botometer	0.0608	*ImpeachTrump*	14	morphonios	27	blackstoneintel.com	5
Retweets RU Prop.	True	Debot	False	*Syria*	13	justinamash	24	aljazeera.com	4

E.4 Chaos (random content) bots

Chaos or *random content* bots retweet and propagate random content. At times these accounts use the random content to hide the purpose of the account. In other words, the signal and message of interest is hidden within the noise of random retweets.

Chaos accounts are identified by content that spans the spectrum of topic and often language. On the next few pages we'll give a few examples of *chaos* accounts. Note that they generally have a uniform distribution over the time of day and have a good distribution of languages.

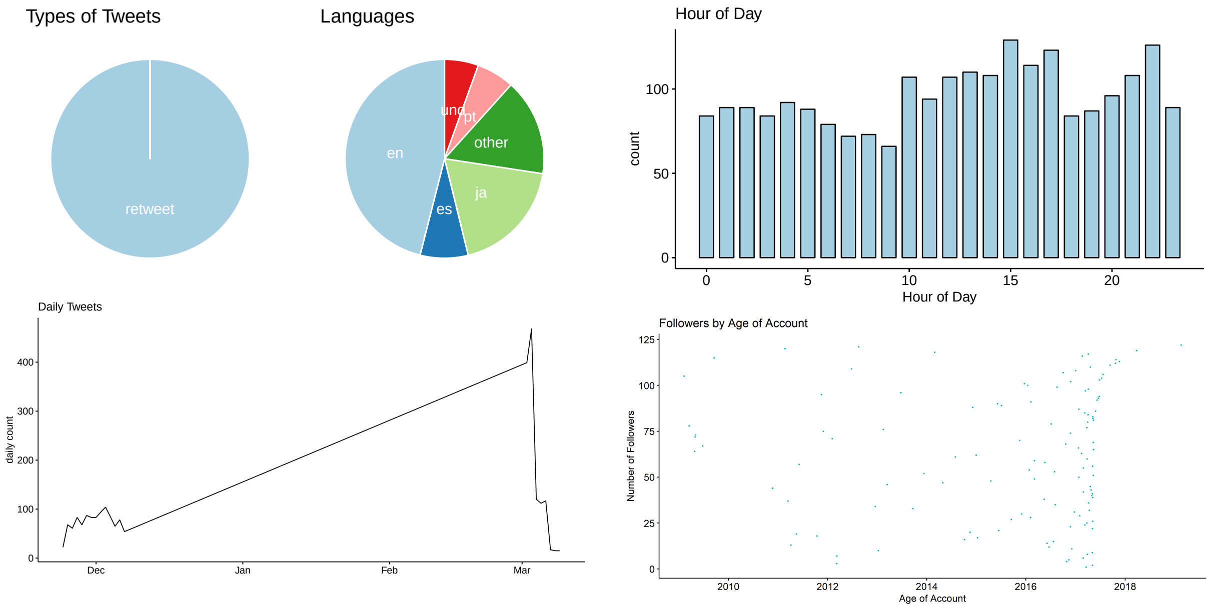


RetweetMaster7
@retweetmaster7

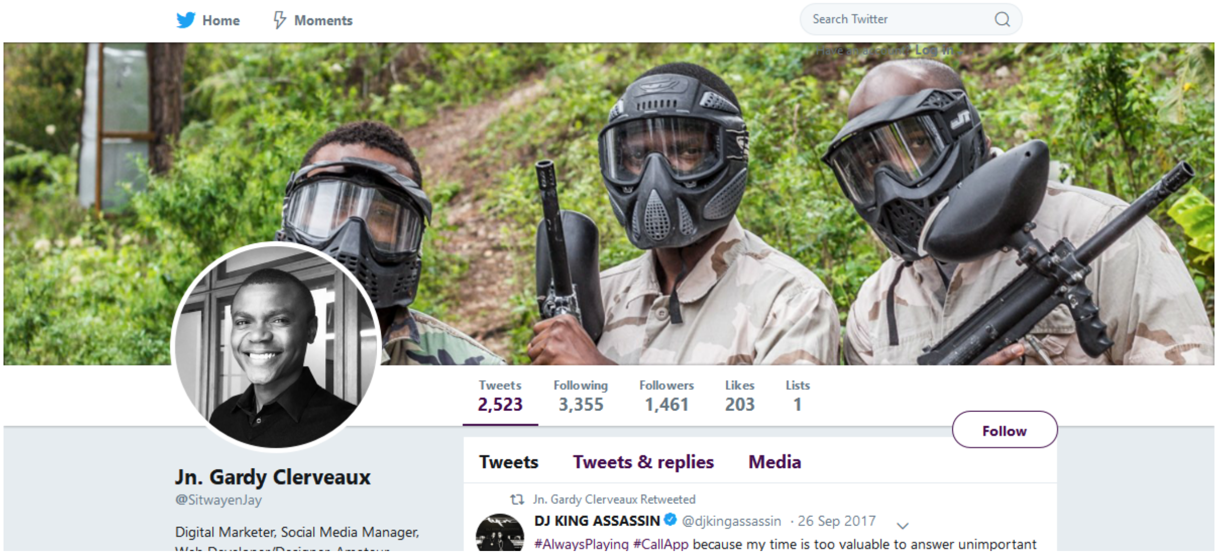
I retweet only most viral content on

RetweetMaster7 Retweeted
ぐらたん @guda12662 · 19 Feb 2018
全部貴女のせい。

U. @rewteetmaster7. We found @retweetmaster7 linked to an attack on NATO. Note the uniform distribution for time of day as well as the wide variety of languages.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.016	bot-hunter Tier 0	False	BTS	38	BTS_twt	51	youtu.be	11
Avg Sentiment (0-1)	0.638	bot-hunter Tier 1	0.3859	iHeartAwards	23	latelateshow	17	goo.gl	10
Dormant Followers	41.0%	bot-hunter Tier 2	0.537	BestFanArmy	18	weareoneEXO	15	bit.ly	8
Bot Followers	27.0%	bot-hunter Tier 3	0.609		16	BT21_	13	tistory.com	4
Avg Mentions	1.14	botometer	0.9206	BTSARMY	12	1theK	13	soompi.com	4
Retweets RU Prop.	False	Debot	False	InfinityWar	11	bts_bighit	10	instagram.com	4



Tweets 2,523 | Following 3,355 | Followers 1,461 | Likes 203 | Lists 1

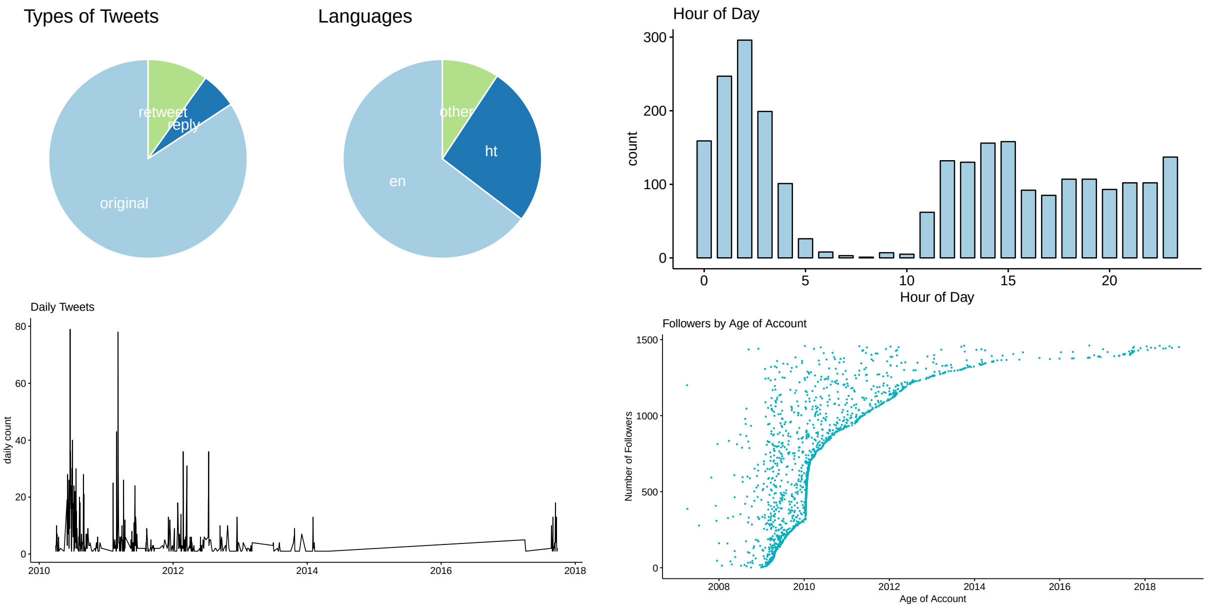
Tweets | Tweets & replies | Media

Jn. Gardy Clerveaux
@SitwayenJay

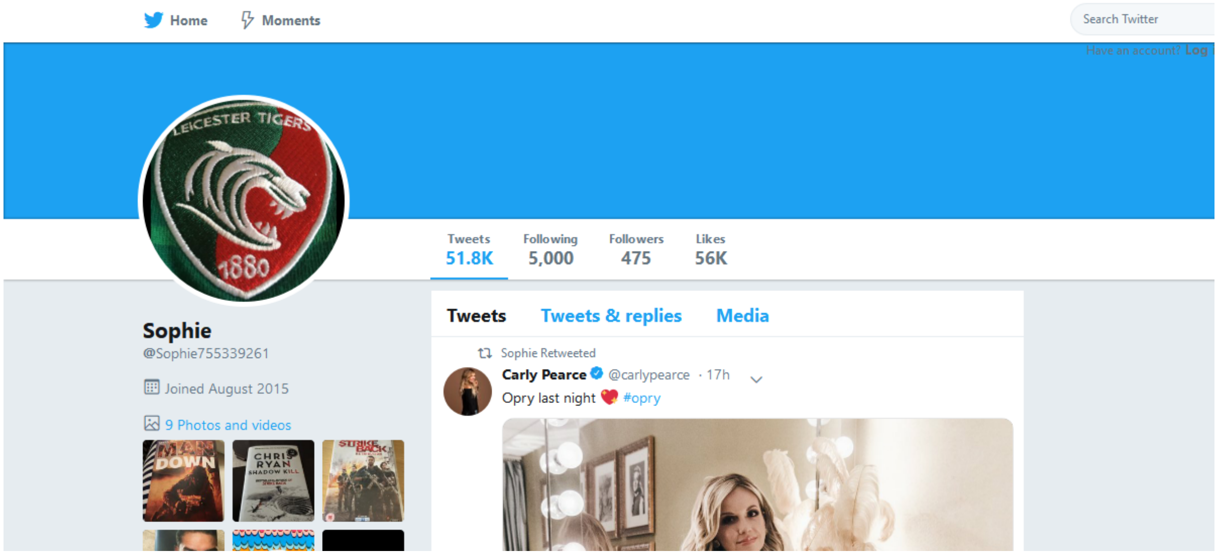
Digital Marketer, Social Media Manager,

Jn. Gardy Clerveaux Retweeted
DJ KING ASSASSIN @djkingassassin · 26 Sep 2017
#AlwaysPlaying #CallApp because my time is too valuable to answer unimportant

V. @SitwayenJay. We found @SitwayenJay linked to an attack on NATO. Note the uniform distribution for time of day as well as the wide variety of languages.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	2.296	bot-hunter Tier 0	False	haiti	29	Giebarreau	180	tmi.me	26
Avg Sentiment (0-1)	0.669	bot-hunter Tier 1	0.5016	Fon_Pale_Mezanmi	28	timoza	124	gamebag.org	18
Dormant Followers	60.3%	bot-hunter Tier 2	0.7255	Haiti	26	carelpedre	71	youtu.be	12
Bot Followers	43.5%	bot-hunter Tier 3	0.560	np	19	craav	54	youtube.com	10
Avg Mentions	1.254	botometer	0.7855	Chanjman	14	DigicelHT	45	bit.ly	9
Retweets RU Prop.	False	Debot	False	oupakagason	12	PaGadAlem	44	ow.ly	9



Home

Moments

Search Twitter

Have an account? Log

Tweets 51.8K

Following 5,000

Followers 475

Likes 56K

Sophie

@Sophie755339261

Joined August 2015

9 Photos and videos

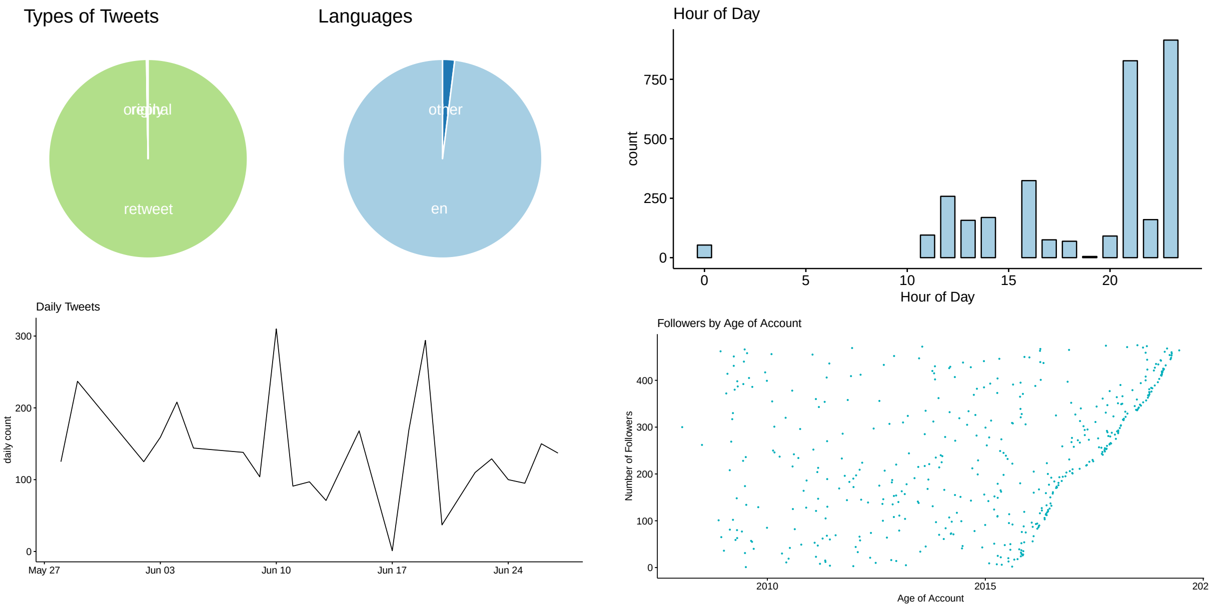
Tweets | Tweets & replies | Media

Sophie Retweeted

Carly Pearce @carlypearce · 17h

Opry last night ❤️ #opry

W. @Sophie755339261. The @Sophie755339261 posts random content in English that seems to hide the more political primary purpose of the account.



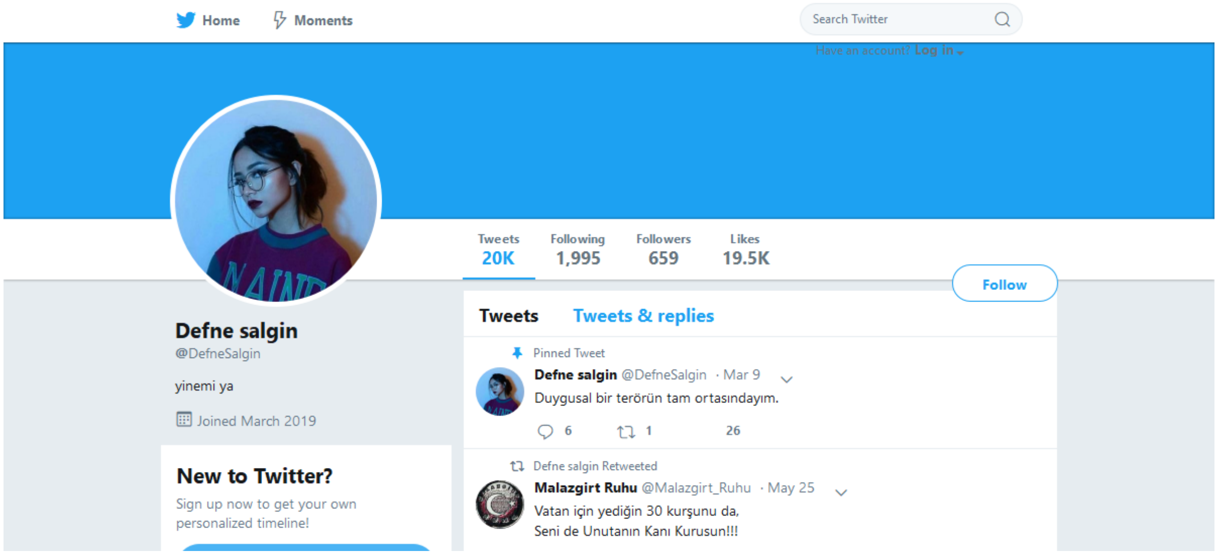
Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	10.526	bot-hunter Tier 0	False	USNavy	17	Fact	764	peoplem.ag	44
Avg Sentiment (0-1)	0.688	bot-hunter Tier 1	0.6417	NYC	16	InspowerMinds	206	trib.al	31
Dormant Followers	41.9%	bot-hunter Tier 2	0.555	SEALTeam	12	Inspire$_US$	170	bit.ly	24
Bot Followers	26.3%	bot-hunter Tier 3	0.642	ChicagoMed	12	MotivatedLiving	155	abcn.ws	8
Avg Mentions	1.203	botometer	0.3366	newyork	9	DavidRoads	109	bbc.in	7
Retweets RU Prop.	False	Debot	False	USStateVisit	8	Quote$_Soup$	94	7ny.tv	7

Coordinated Bots

Coordinated or *correlated* bots are two or more bots that are all cloned to send roughly the same tweets/retweets at roughly the same time. When you look at the statistics of correlated accounts, you will see that the content and timing are very similar if not exact. Note that this correlation is created in one of two ways. Either 1) both accounts are created by the same bot "puppet-master" and were intentionally cloned, or a bot-puppet master created an account that exactly replicates another account that he/she does not own.

Several algorithms have been developed to help researchers and analysts identify correlated accounts. Notable among these methods the calculation of warped correlation, which powers the underlying algorithm of the *Debot* unsupervised bot detection algorithm.

On the next two pages we give an example of two accounts that are exactly correlated. Flipping back and forth you will notice that all distributions are exactly mirrored for both accounts.



Defne salgin
@DefneSalgin

yinemi ya

Joined March 2019

New to Twitter?

Sign up now to get your own personalized timeline!

Pinned Tweet

Defne salgin @DefneSalgin · Mar 9

Duygusal bir terörün tam ortasındayım.

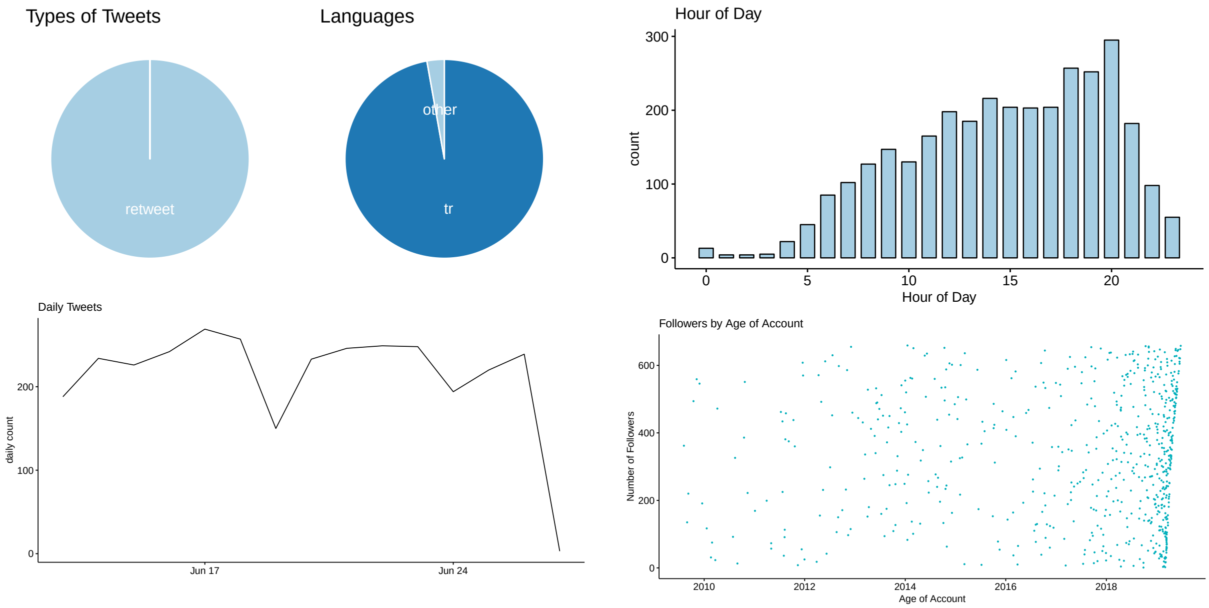
6 1 26

Defne salgin Retweeted

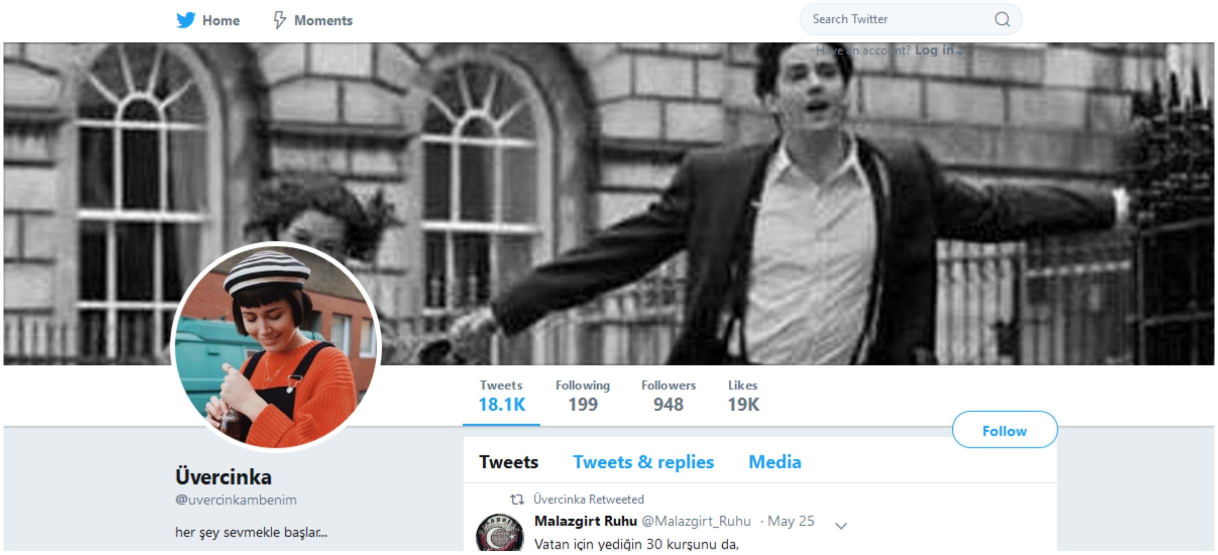
Malazgirt Ruhu @Malazgirt_Ruhu · May 25

Vatan için yediğin 30 kurşunu da,
Seni de Unutanın Kanı Kurusun!!!

X. @DefneSaligin. Compare this page with the following page...they have exactly the same plots.



Metric	Value	Model	Probability	hashtag	count	Mention	Count	domain	count
Frd/Fol Ratio	3.027	bot-hunter Tier 0	False	BYvsEİ	13	TurkiyeHareket	139	bit.ly	67
Avg Sentiment (0-1)	0.702	bot-hunter Tier 1	0.5173	herşeyçokgüzelolacak	13	drekremculfa	105	hepsiburada.com	31
Dormant Followers	2.1%	bot-hunter Tier 2	0.5787	ekremimamoglu	13	ahmetbozkurt140	93	Sistemkoin.com	15
Bot Followers	12.0%	bot-hunter Tier 3	0.606	matkitap	12	gazeteistiklal	92	natrumin.com.tr	10
Avg Mentions	1.191	botometer	0.6164	Natrumin	12	LokmanCagirici	89	kevserantik.com	9
Retweets RU Prop.	False	Debot	True	imamoglu	11	sevkodiyosun	78	n11.com	9



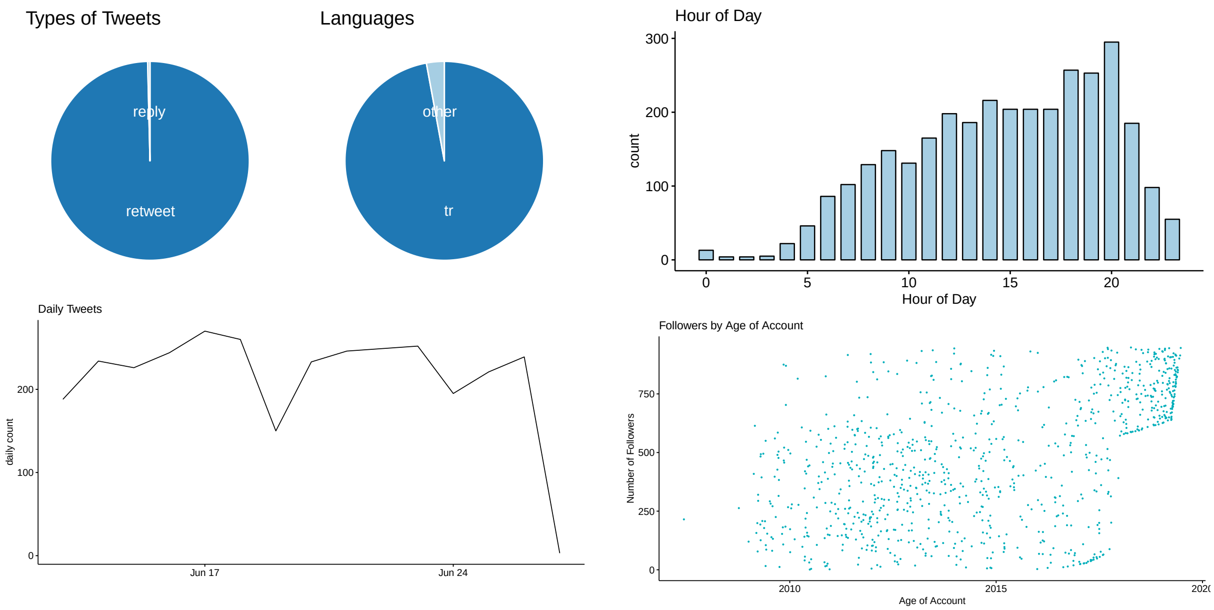
Tweets 18.1K · Following 199 · Followers 948 · Likes 19K

Üvercinka
@uvercinkambenim

her şey sevmekle başlar...

Üvercinka Retweeted
Malazgirt Ruhu @Malazgirt_Ruhu · May 25
Vatan için yediğin 30 kurşunu da,

Y. @uvercinkambenim. Compare this page with the previous page...they have exactly the same plots.

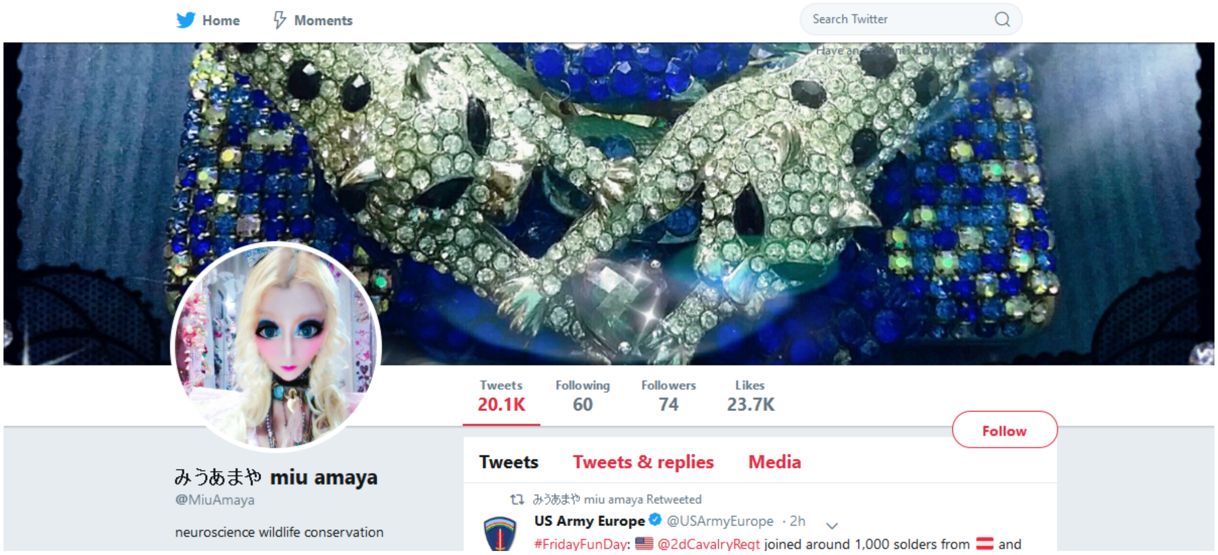


Metric	Value	Model	Probability	hashtag	count	Mention	Count	domain	count
Frd/Fol Ratio	0.21	bot-hunter Tier 0	False	BYvsEİ	13	TurkiyeHareket	139	bit.ly	67
Avg Sentiment (0-1)	0.702	bot-hunter Tier 1	0.4399	ekremimamoglu	13	drekremculfa	105	hepsiburada.com	31
Dormant Followers	43.6%	bot-hunter Tier 2	0.5589	herşeyçokgüzelolacak	13	ahmetbozkurt140	93	Sistemkoin.com	15
Bot Followers	57.0%	bot-hunter Tier 3	0.691	Natrumin	12	gazeteistiklal	92	natrumin.com.tr	10
Avg Mentions	1.192	botometer	0.4337	matkitap	12	LokmanCagirici	89	n11.com	9
Retweets RU Prop.	False	Debot	True	imamoglu	11	sevkodiyosun	78	kevserantik.com	9

E.5 Social Influence Bots

Social Influence bots attempt to manipulate the network and algorithms of social media by intentionally mentioning, retweeting, and liking each other. These efforts increase the visibility of these respective accounts and make them look more popular than they really are. Additionally, these efforts manipulate the algorithms that prioritize content.

At times these social influence bots and their respective bot-nets can be identified by identifying dense sub-graphs in the conversational network.

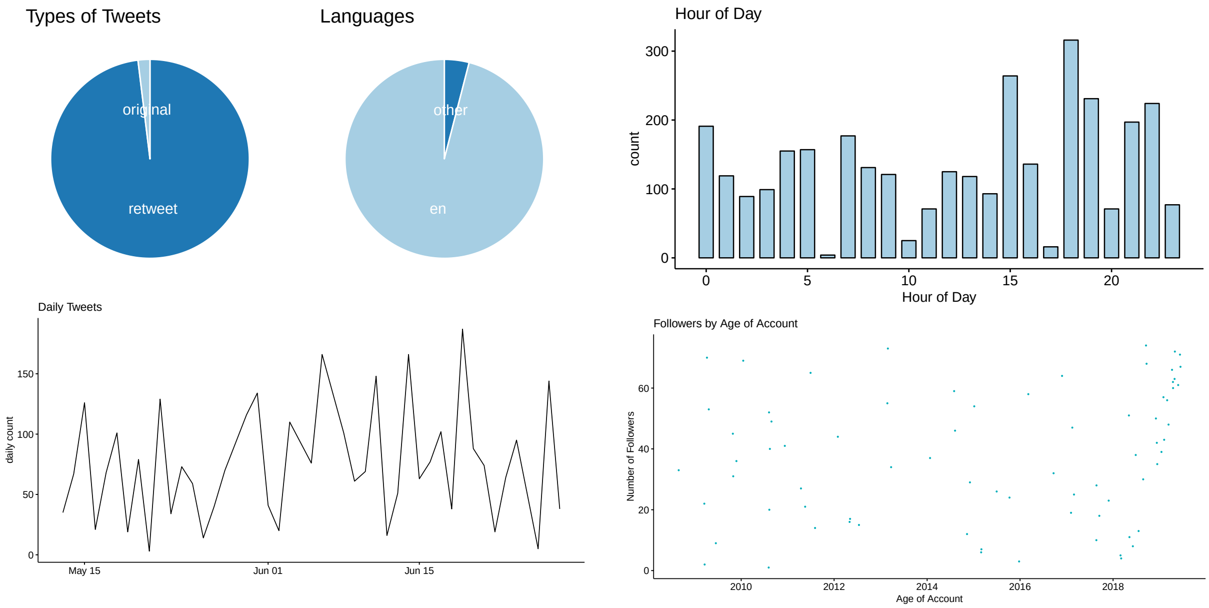


みうあまや miu amaya
@MiuAmaya

neuroscience wildlife conservation

みうあまや miu amaya Retweeted
US Army Europe @USArmyEurope · 2h
#FridayFunDay: 🇺🇸 @2dCavalryRegt joined around 1,000 solders from 🇦🇹 and

Z. @MiuAmaya. Japanese anime bot that seems to tweet about NATO often.

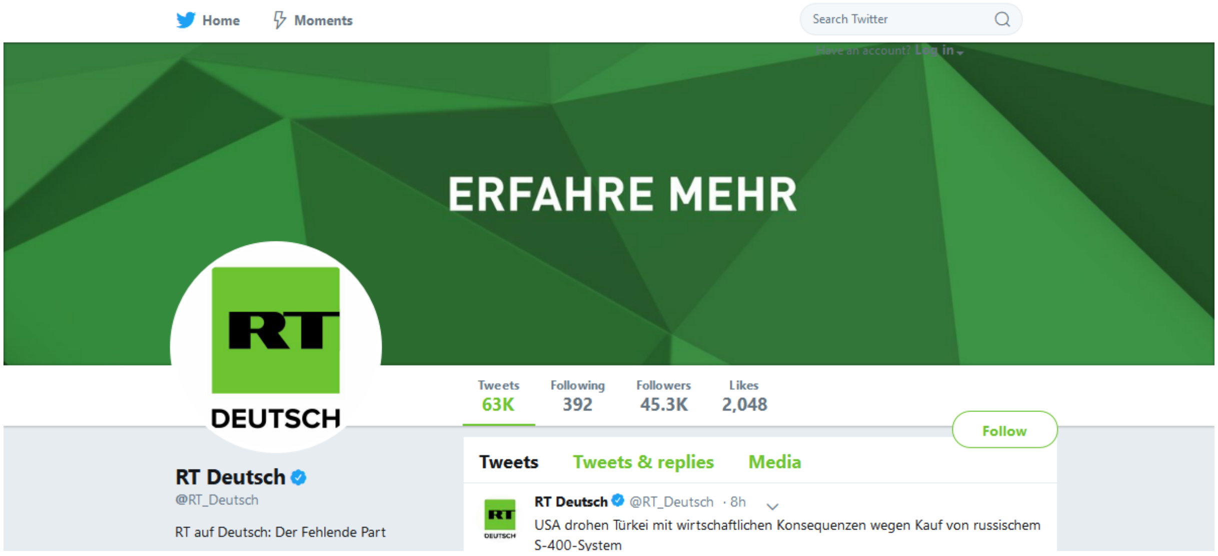


Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.811	bot-hunter Tier 0	False	NATO	124	starsandstripes	145	wapo.st	68
Avg Sentiment (0-1)	0.707	bot-hunter Tier 1	0.3897	NIH	54	YonkersMBK	124	bit.ly	49
Dormant Followers	9.5%	bot-hunter Tier 2	0.4452	NYSMBK	36	PointsofLight	111	dlvr.it	41
Bot Followers	20.3%	bot-hunter Tier 3	0.717	DDay75	35	NIH	101	youtu.be	26
Avg Mentions	1.463	botometer	0.6164	WeAreMBK	33	JoeBiden	101	stripes.com	19
Retweets RU Prop.	True	Debot	False	HIV	21	washingtonpost	97	nyti.ms	13

E.6 News Bot Accounts

There are thousands of beneficial or productive bots that exist. One common type of productive bot is the news bot. A news bot can be used by both news producing agencies and news aggregators. Many news producing agencies have bot accounts that automatically post summaries and links to their news articles. Other news aggregating bot accounts monitor multiple news sites and automatically retweet or summarize news from a myriad of sites.

Note that some accounts that are set up to look like news bots have malicious intent. For example, Russia's Internet Research Agency (IRA) used accounts that looked like city news aggregators in the runnup to the 2016 US elections. These accounts had screen names like *NewOrleansON* and *KansasDailyNews*.



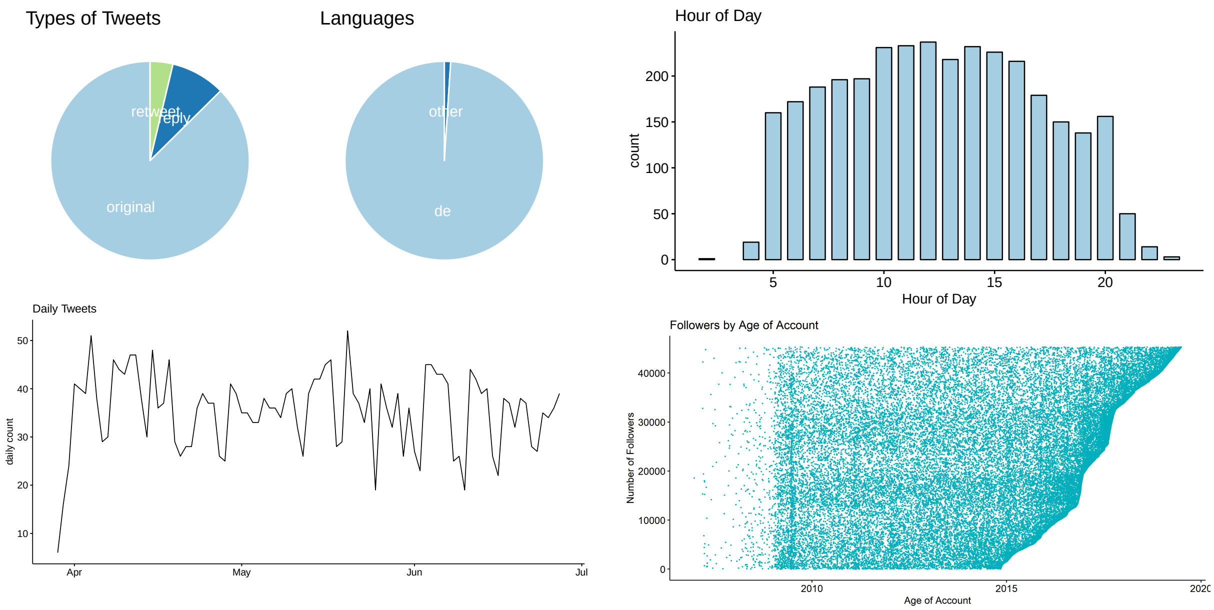
RT Deutsch @RT_Deutsch

RT auf Deutsch: Der Fehlende Part

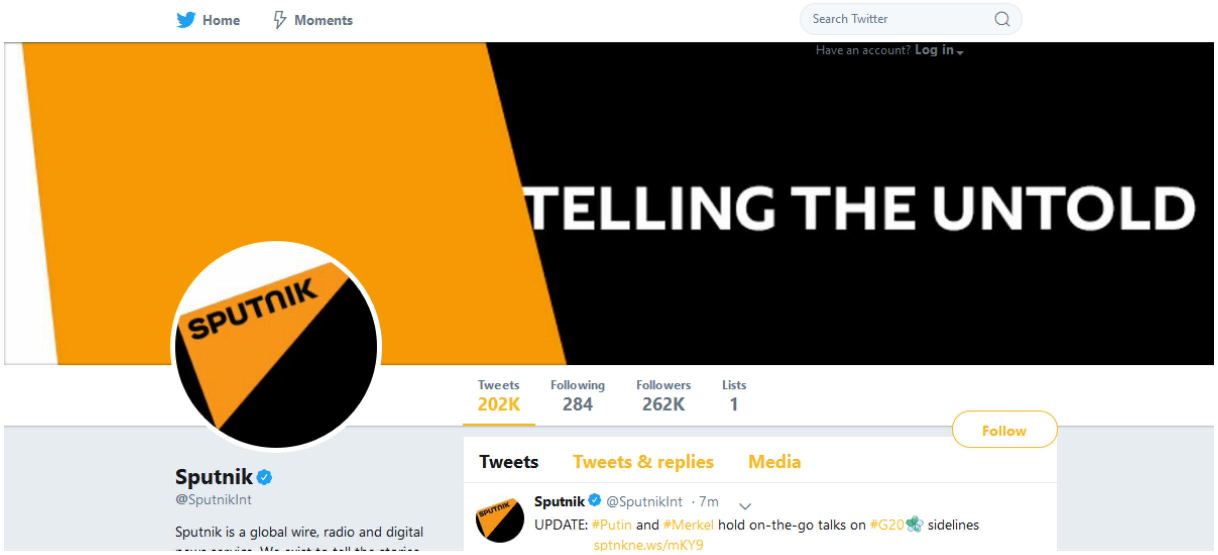
RT Deutsch @RT_Deutsch · 8h

USA drohen Türkei mit wirtschaftlichen Konsequenzen wegen Kauf von russischem S-400-System

. **@RT_Deutsch.** This shows a Russian News Company propagating their news in Germany.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.009	bot-hunter Tier 0	False	USA	83	451_grad	78	rt.com	2343
Avg Sentiment (0-1)	0.588	bot-hunter Tier 1	0.4859	LIVE	70	RT_Deutsch	40	youtube.com	121
Dormant Followers	44.4%	bot-hunter Tier 2	0.3103	Assange	67	FWarweg	16	youtu.be	68
Bot Followers	52.5%	bot-hunter Tier 3	0.429	Venezuela	66	BILD	9	rbth.com	10
Avg Mentions	1.368	botometer	0.1112	Russland	61	JulianRoepcke	9	facebook.com	4
Retweets RU Prop.	True	Debot	False	Iran	53	GertEwen	9	451grad.com	3



Home Moments

Search Twitter

Have an account? Log in

TELLING THE UNTOLD

SPUTNIK

Tweets	Following	Followers	Lists
202K	284	262K	1

Follow

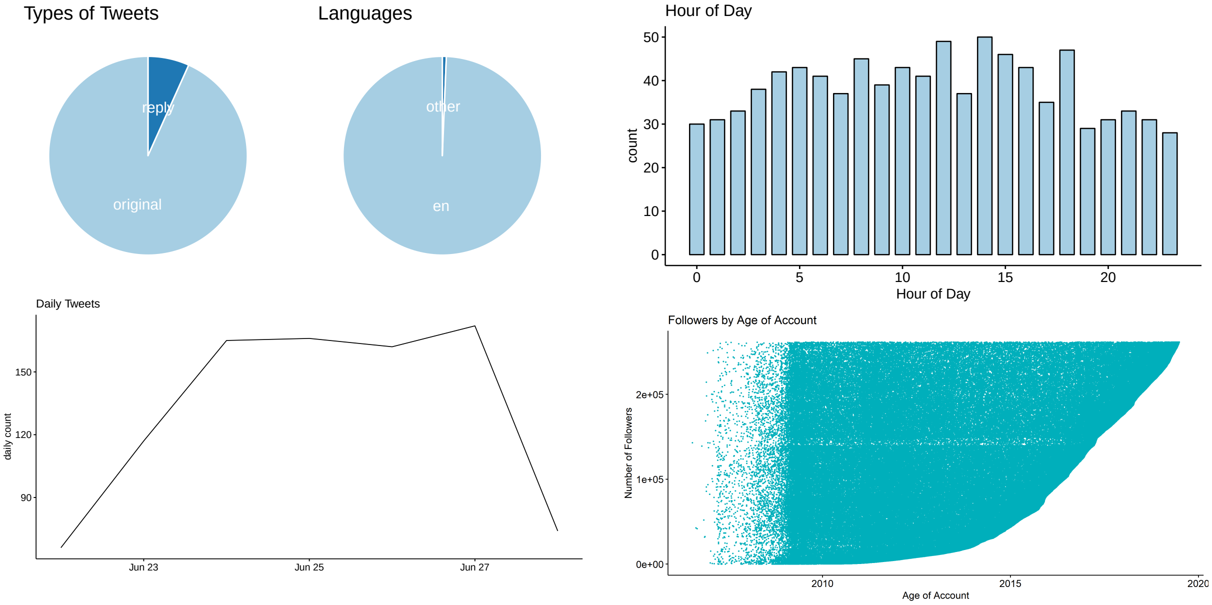
Sputnik
@SputnikInt

Sputnik is a global wire, radio and digital

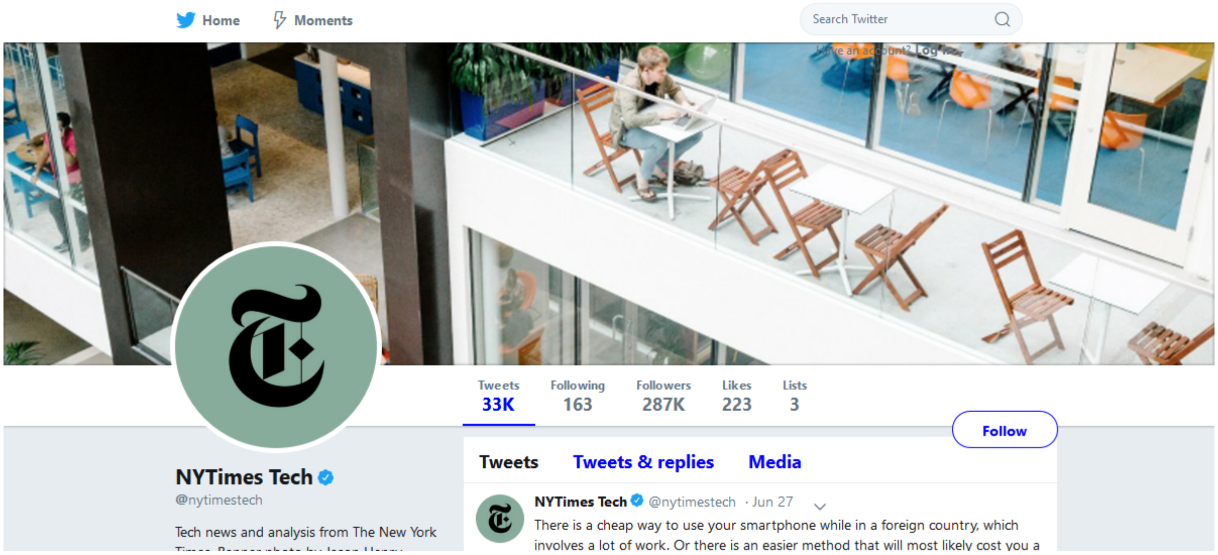
Tweets Tweets & replies Media

Sputnik @SputnikInt · 7m
UPDATE: #Putin and #Merkel hold on-the-go talks on #G20 sidelines
sptnkne.ws/mKY9

. **@SputnikInt.** This shows a Russian state sponsored news feed.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.001	bot-hunter Tier 0	False	Iran	73	realDonaldTrump	119	sptnkne.ws	685.0
Avg Sentiment (0-1)	0.48	bot-hunter Tier 1	0.1732	Russia	33	Huawei	21		
Dormant Followers	37.1%	bot-hunter Tier 2	0.4922	India	29	RTErdogan	19		
Bot Followers	48.4%	bot-hunter Tier 3	0.418	China	27	NATO	12		
Avg Mentions	1.367	botometer	0.0562	Putin	23	BorisJohnson	12		
Retweets RU Prop.	True	Debot	False	Turkey	16	JZarif	10		



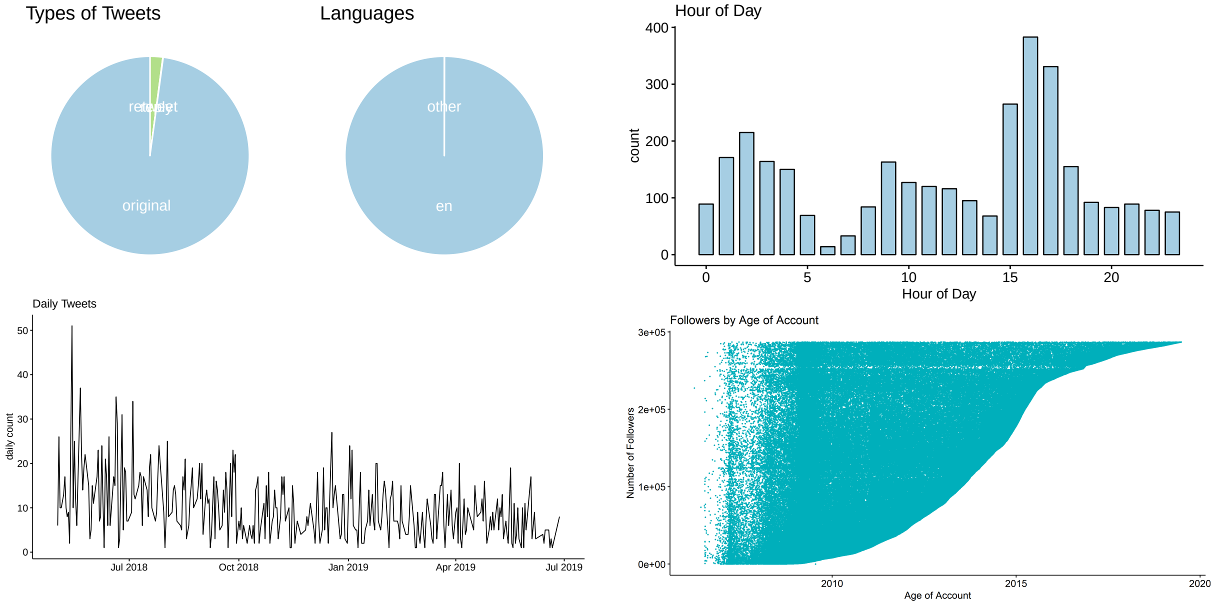
NYTimes Tech
@nytimestech

Tech news and analysis from The New York Times. Banner photo by Jason Henry

NYTimes Tech @nytimestech · Jun 27

There is a cheap way to use your smartphone while in a foreign country, which involves a lot of work. Or there is an easier method that will most likely cost you a

. **@nytimestech.** This shows how an American news company can use Twitter bots and Cyborgs to propagate their news.

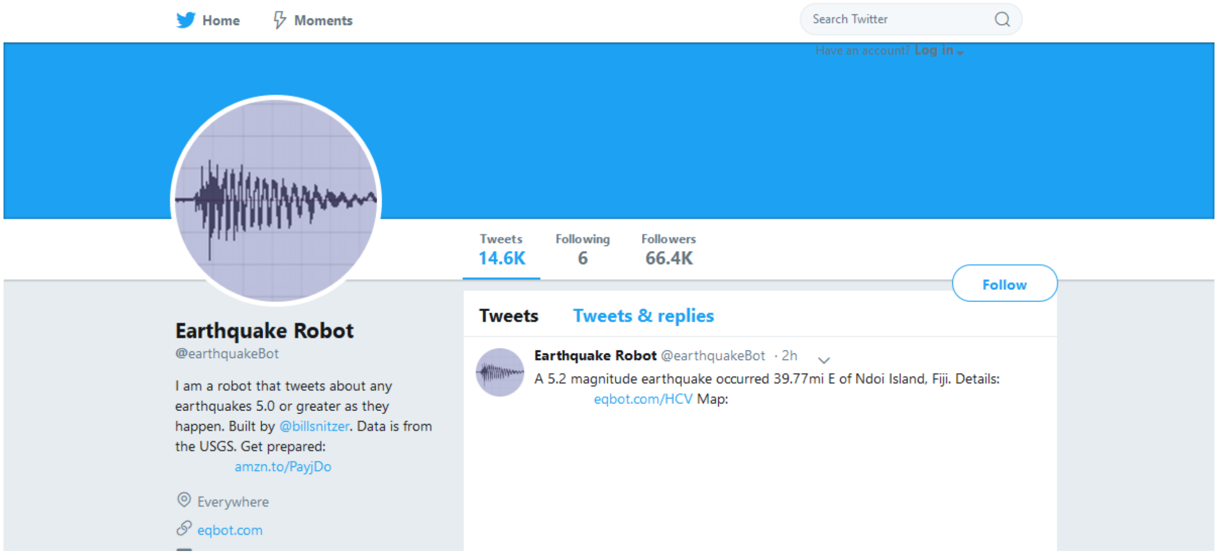


Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.001	bot-hunter Tier 0	False	nevertweet	1.0	bxchen	41	nyti.ms	607
Avg Sentiment (0-1)	0.549	bot-hunter Tier 1	0.2061	TimesTalks	1.0	fmanjoo	26	JD.com	4
Dormant Followers	53.1%	bot-hunter Tier 2	0.377	MeToo	1.0	kevinroose	23	nytimes.com	4
Bot Followers	53.7%	bot-hunter Tier 3	0.434			jacknicas	14	Drive.ai	4
Avg Mentions	1.222	botometer	0.0299			SteveLohr	12	WordPress.com	1
Retweets RU Prop.	False	Debot	False			MikeIsaac	10	trib.al	1

E.7 Overt Bot Accounts

Overt bots are accounts that explicitly claim to be a bot. Many of these accounts are either fun hobby bots or are beneficial and productive bots. Some of these, however, can be propaganda bots. By explicitly claiming to be a bot in the description, these accounts can be allowed to generate bot like activity on Twitter. But remember, the account description does not travel with the bot tweets.

The next two pages have some examples of overt bots, some of them beneficial, others less so.



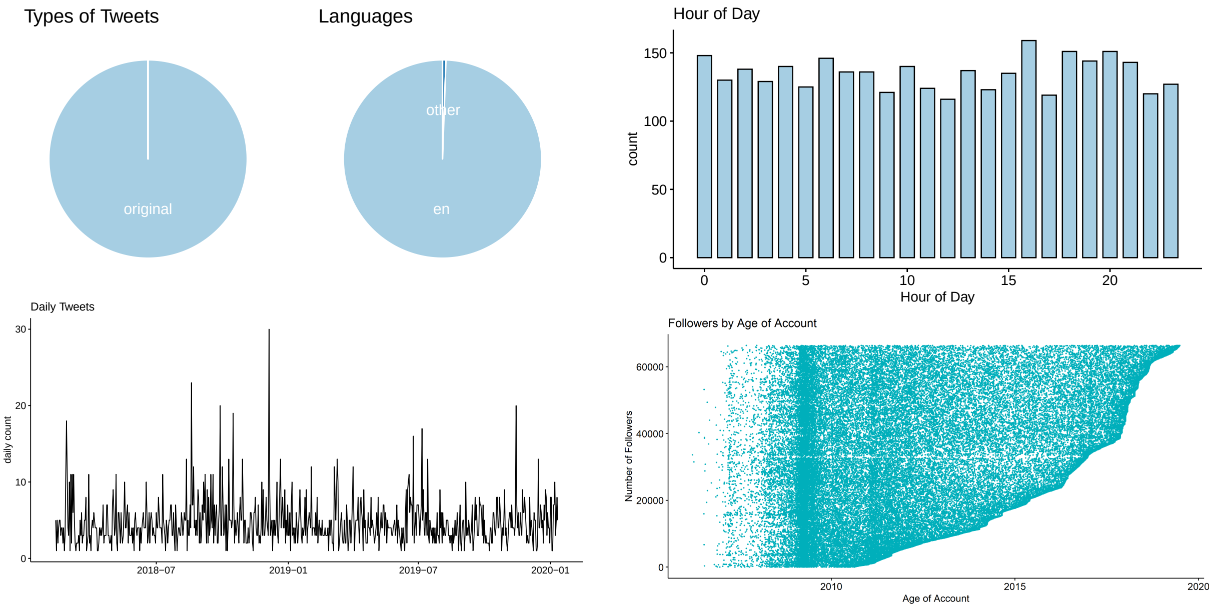
Earthquake Robot
@earthquakeBot

I am a robot that tweets about any earthquakes 5.0 or greater as they happen. Built by @billsnitzer. Data is from the USGS. Get prepared: amzn.to/PayjDo

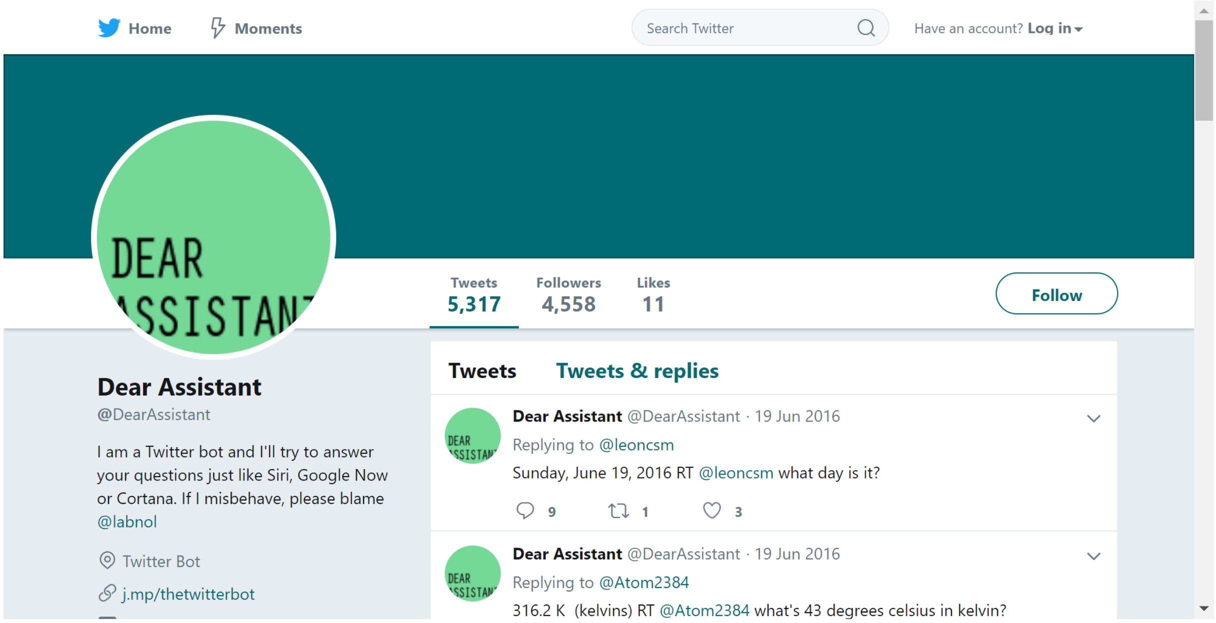
Everywhere
eqbot.com

Earthquake Robot @earthquakeBot · 2h
A 5.2 magnitude earthquake occurred 39.77mi E of Ndoi Island, Fiji. Details: eqbot.com/HCV Map:

. **@earthquakeBot.** This account is a very beneficial account that notifies people of earthquakes.



Metric	Value	Model	Probability	Count	Hashtag	Count	Mention	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False					eqbot.com	6385.0
Avg Sentiment (0-1)	0.571	bot-hunter Tier 1	0.3145						
Dormant Followers	29.7%	bot-hunter Tier 2	0.2063						
Bot Followers	46.7%	bot-hunter Tier 3	0.315						
Avg Mentions		botometer	0.3552						
Retweets RU Prop.	False	Debot	False						



Tweets **Tweets & replies**

Dear Assistant
@DearAssistant

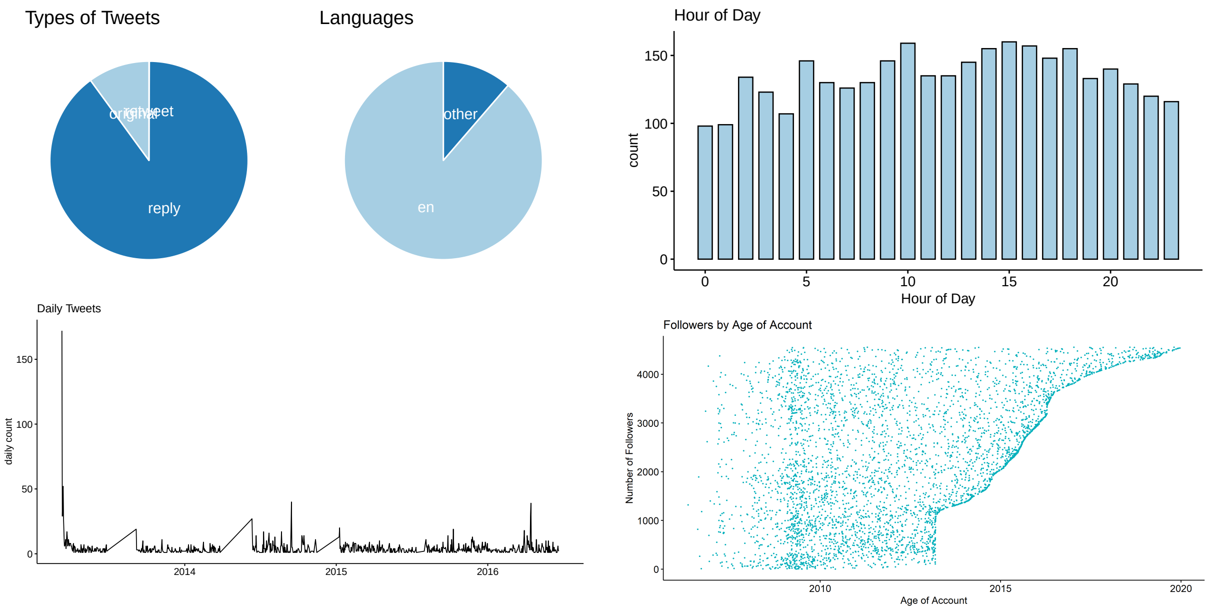
I am a Twitter bot and I'll try to answer your questions just like Siri, Google Now or Cortana. If I misbehave, please blame @labnol

Twitter Bot
j.mp/thetwitterbot

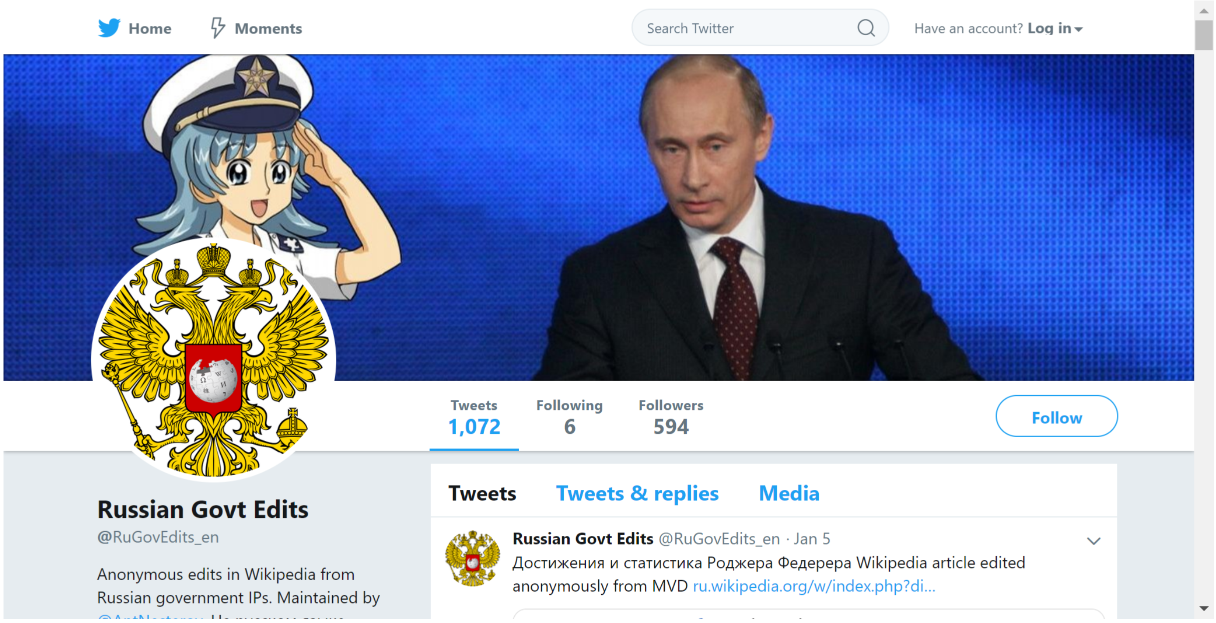
Dear Assistant @DearAssistant · 19 Jun 2016
Replying to @leoncsm
Sunday, June 19, 2016 RT @leoncsm what day is it?
9 1 3

Dear Assistant @DearAssistant · 19 Jun 2016
Replying to @Atom2384
316.2 K (kelvins) RT @Atom2384 what's 43 degrees celsius in kelvin?

. **@DearAssistant.** @DearAssistant provides an example of a personal assistant bot. This bot answers questions in much the same way that Apple Siri or Microsoft Cortana is able to answer questions.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False	*eMOOCs2015*	1.0	labnol	162	i.am	1.0
Avg Sentiment (0-1)		bot-hunter Tier 1	0.3988	*0000ff*	1.0	howto_guides	32	t.co	1.0
Dormant Followers	40.2%	bot-hunter Tier 2	0.2838	*INDvsAUS*	1.0	Atom2384	30	talltweets.com	1.0
Bot Followers	39.9%	bot-hunter Tier 3	0.315	*FF00FF*	1.0	howto	29	java.com	1.0
Avg Mentions	1.07	botometer	0.3552			FabioReyes16	27		
Retweets RU Prop.	False	Debot	False			ktmweather	23		

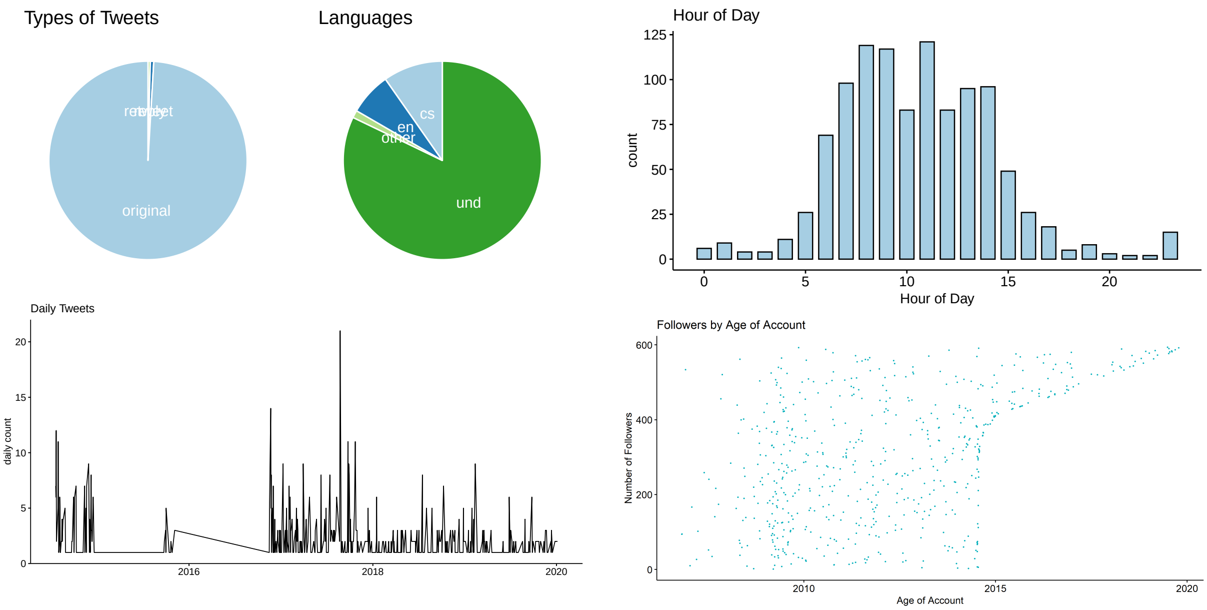


Russian Govt Edits
@RuGovEdits_en

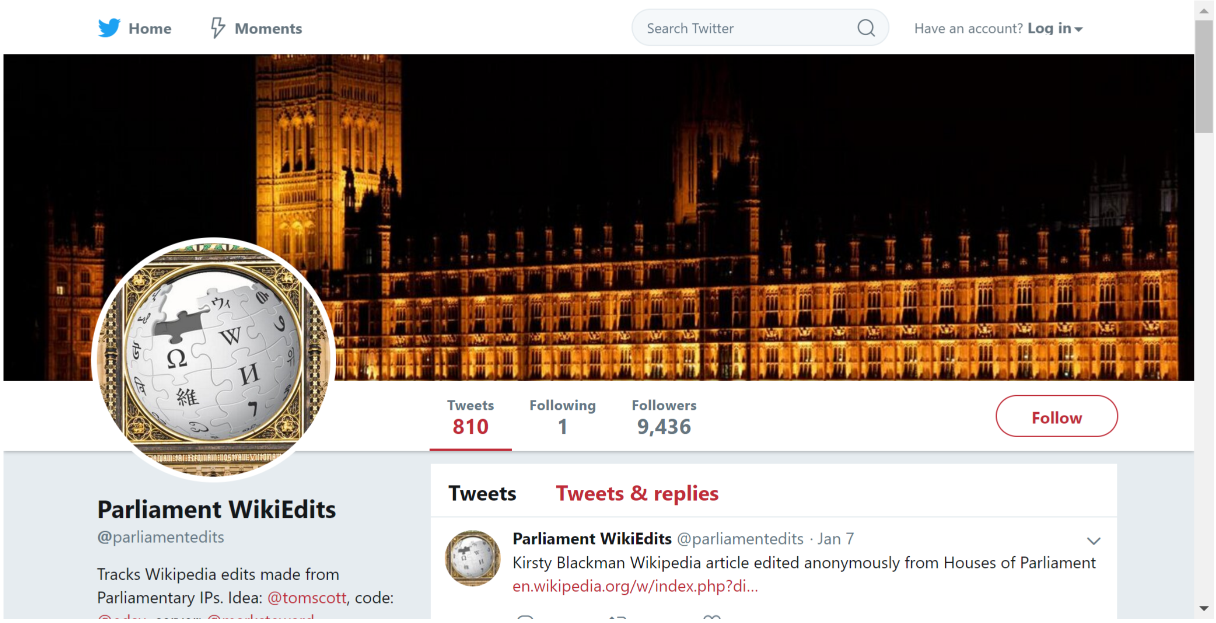
Anonymous edits in Wikipedia from Russian government IPs. Maintained by

Russian Govt Edits @RuGovEdits_en · Jan 5
Достижения и статистика Роджера Федерера Wikipedia article edited anonymously from MVD ru.wikipedia.org/w/index.php?di…

. **@RuGovEdits_en.** @RuGovEdits_en is an overt bot that notifies followers of any Wikipedia edits from Russian Government IP space.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.01	bot-hunter Tier 0	False			RuGovEdits	4	wikipedia.org	835
Avg Sentiment (0-1)		bot-hunter Tier 1	0.1928			AntNesterov	2	wikidata.org	12
Dormant Followers	30.6%	bot-hunter Tier 2	0.3029			trsonis	1	github.io	4
Bot Followers	42.8%	bot-hunter Tier 3	0.324			RuBlacklist_en	1	calvertjournal.com	1
Avg Mentions	1.583	botometer	0.5362			nysthee	1	wikimedia.org	1
Retweets RU Prop.	False	Debot	False			katewalters	1	globalvoicesonline.org	1



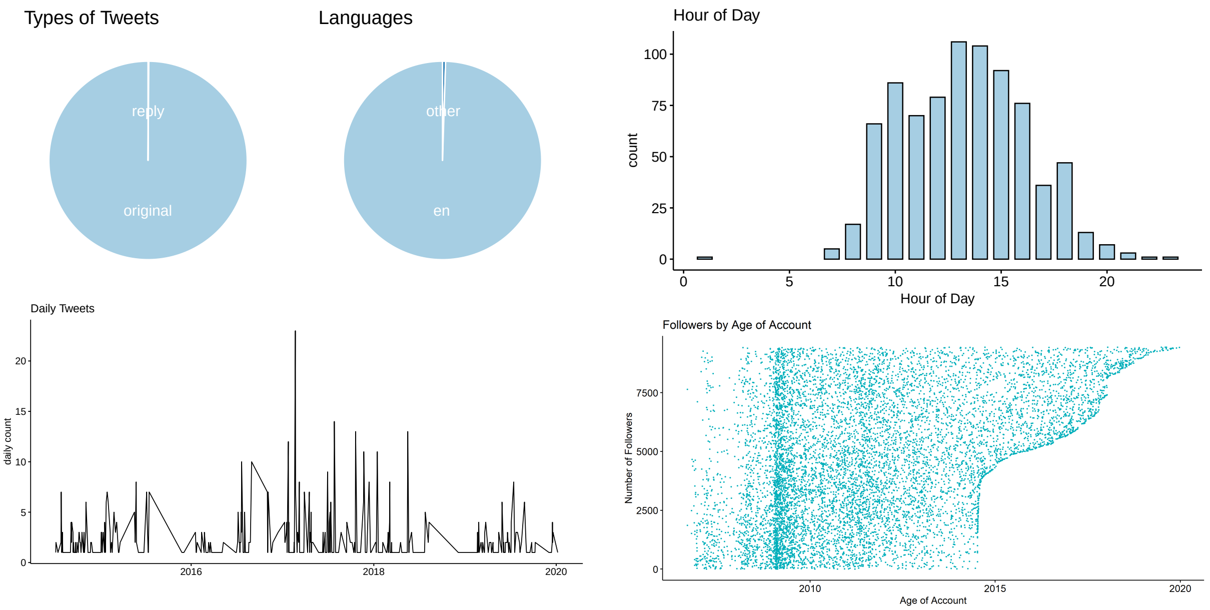
Parliament WikiEdits
@parliamentedits

Tracks Wikipedia edits made from Parliamentary IPs. Idea: @tomscott, code:

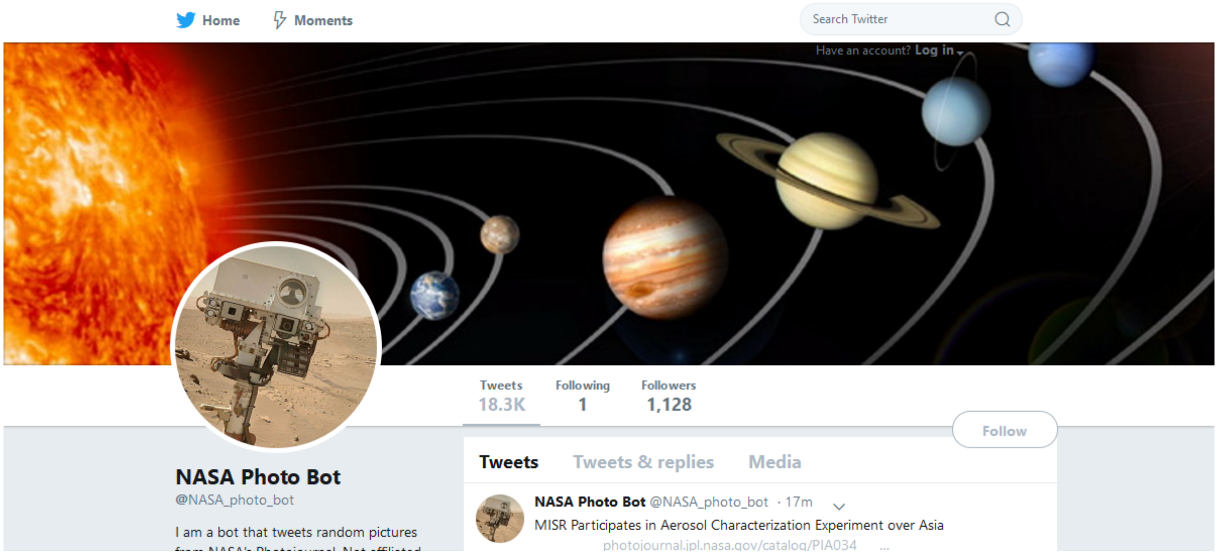
Tweets Tweets & replies

Parliament WikiEdits @parliamentedits · Jan 7
Kirsty Blackman Wikipedia article edited anonymously from Houses of Parliament
en.wikipedia.org/w/index.php?di...

. **@parliamentedits.** @parliamentedits is an overt bot that notifies followers of any changes to Wikipedia made from the IP Space associated with the British Parliament.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False			edsu	2	wikipedia.org	795.0
Avg Sentiment (0-1)		bot-hunter Tier 1	0.295			jonty	1	github.com	2.0
Dormant Followers	17.4%	bot-hunter Tier 2	0.2591			congressedits	1	wikimedia.org	2.0
Bot Followers	36.3%	bot-hunter Tier 3	0.347			RiksdagWikiEdit	1	whatdotheyknow.com	2.0
Avg Mentions	1.8	botometer	0.4136			parliamentedits	1		
Retweets RU Prop.	False	Debot	False			tomscott	1		

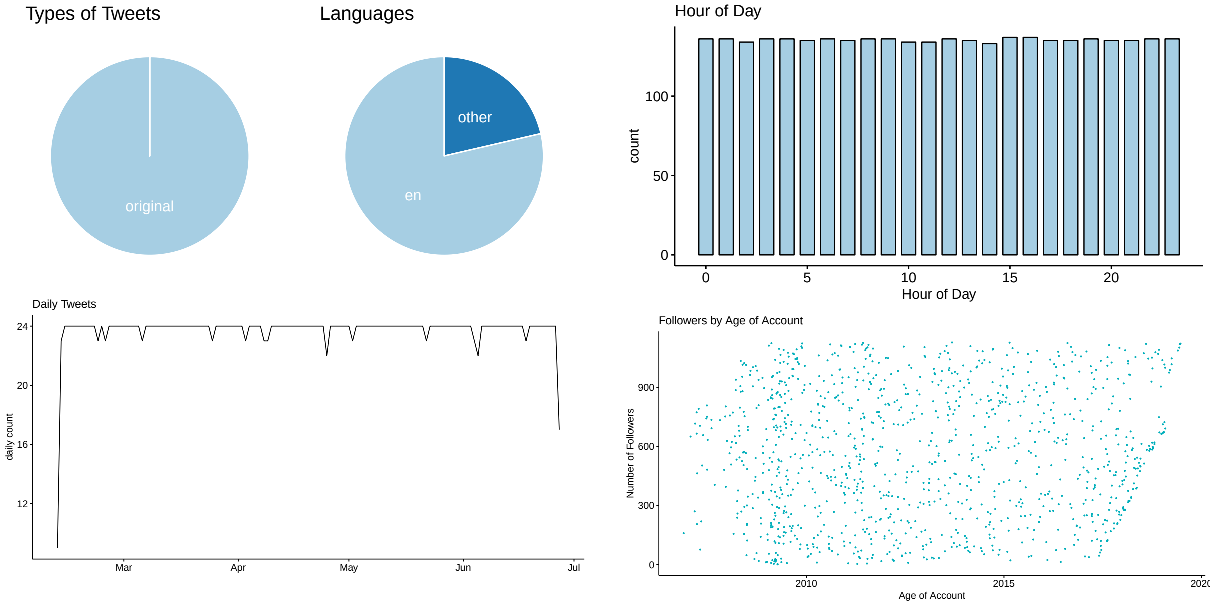


NASA Photo Bot
@NASA_photo_bot

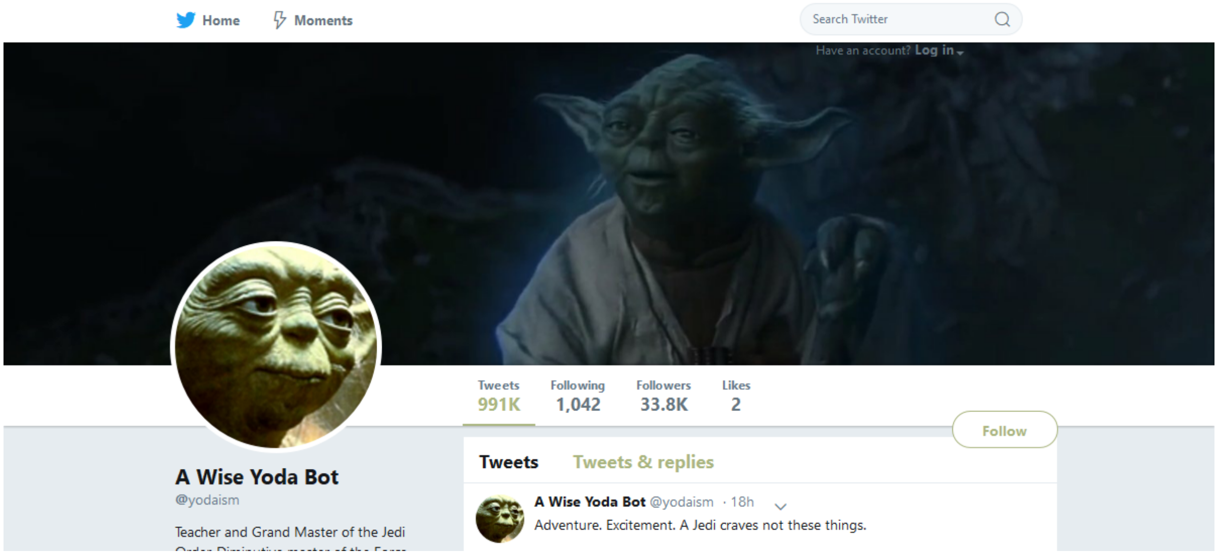
I am a bot that tweets random pictures
from NASA's Photojournal. Not affiliated

NASA Photo Bot @NASA_photo_bot · 17m
MISR Participates in Aerosol Characterization Experiment over Asia
photojournal.jpl.nasa.gov/catalog/PIA034 ...

. **@NASA**$_{p}hoto_{b}ot$. This is another positive bot that posts NASA pictures.



Metric	Value	Model	Probability	Count	Hashtag	Count	Mention	Domain	Count
Frd/Fol Ratio	0.001	bot-hunter Tier 0	False					nasa.gov	3248.0
Avg Sentiment (0-1)	0.629	bot-hunter Tier 1	0.225						
Dormant Followers	10.3%	bot-hunter Tier 2	0.0907						
Bot Followers	40.7%	bot-hunter Tier 3	0.389						
Avg Mentions		botometer	0.4745						
Retweets RU Prop.	False	Debot	False						



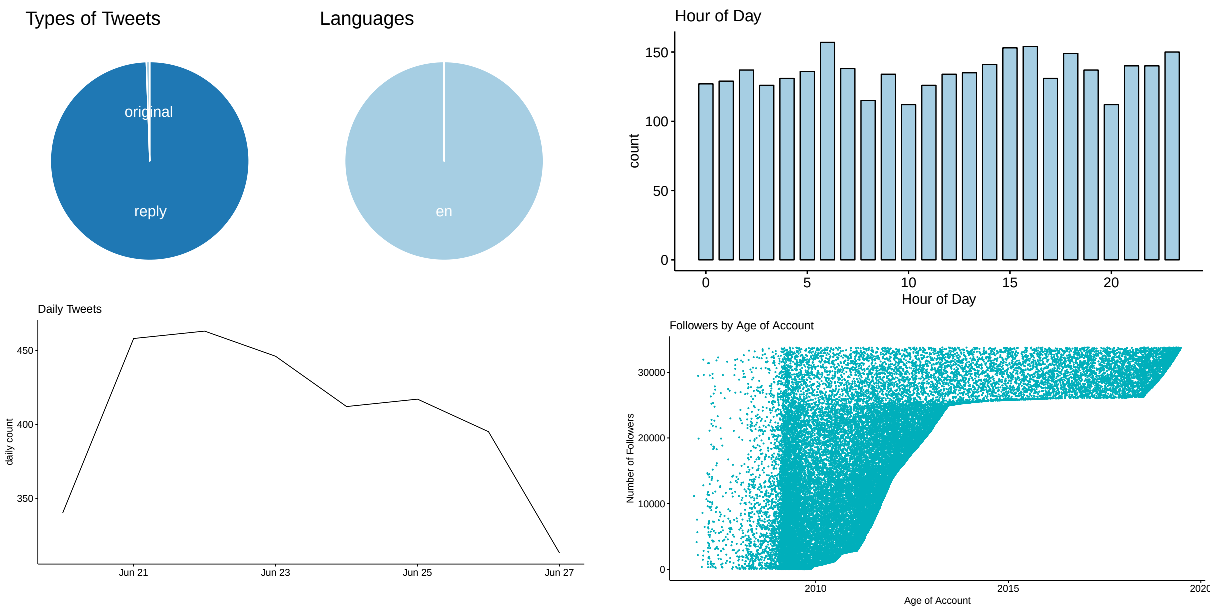
Tweets 991K Following 1,042 Followers 33.8K Likes 2

A Wise Yoda Bot
@yodaism

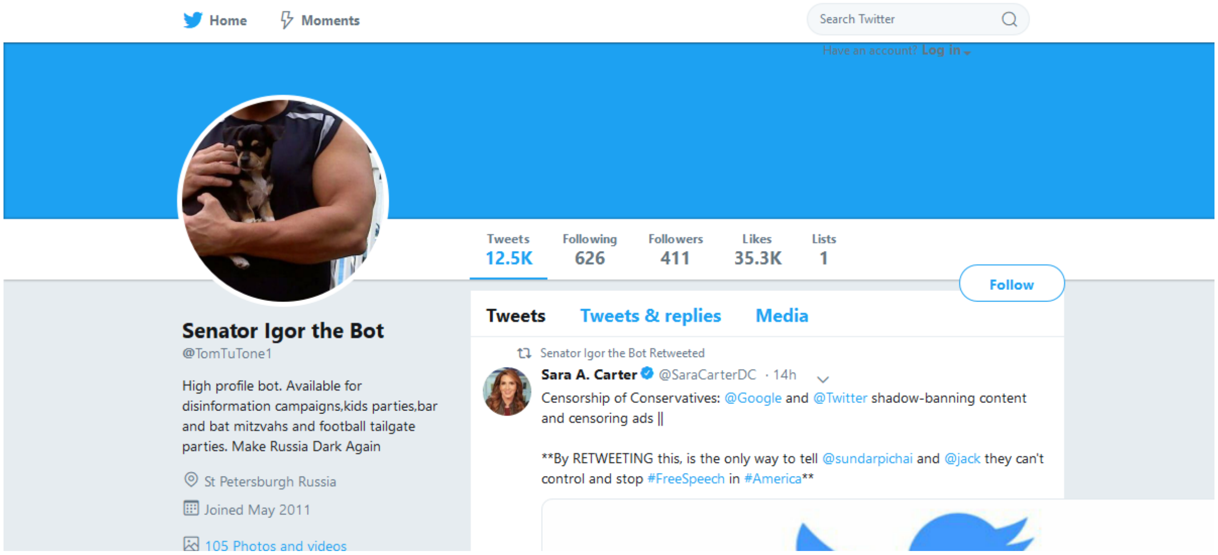
Teacher and Grand Master of the Jedi
Order Diminutive master of the Force

A Wise Yoda Bot @yodaism · 18h
Adventure. Excitement. A Jedi craves not these things.

. **@yodaism.** Fun hobby bot for all the Star Wars fans out there.



Metric	Value	Model	Probability	Count	Hashtag	Mention	Count	Count	Domain
Frd/Fol Ratio	0.031	bot-hunter Tier 0	False			StarWarsInforms	31		
Avg Sentiment (0-1)	0.736	bot-hunter Tier 1	0.3303			NickFoxFighter	13		
Dormant Followers	38.2%	bot-hunter Tier 2	0.1718			wcn483	13		
Bot Followers	25.1%	bot-hunter Tier 3	0.339			442ndSiege	12		
Avg Mentions	1	botometer	0.1947			Salt_SmashBook	10		
Retweets RU Prop.	False	Debot	False			HeckingHeckMe	10		



Senator Igor the Bot
@TomTuTone1

High profile bot. Available for disinformation campaigns,kids parties,bar and bat mitzvahs and football tailgate parties. Make Russia Dark Again

St Petersburgh Russia

Joined May 2011

105 Photos and videos

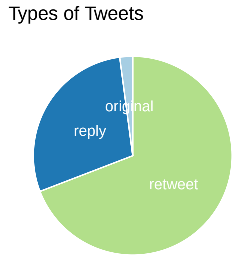
Senator Igor the Bot Retweeted

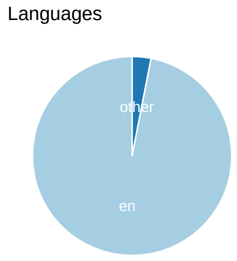
Sara A. Carter @SaraCarterDC · 14h

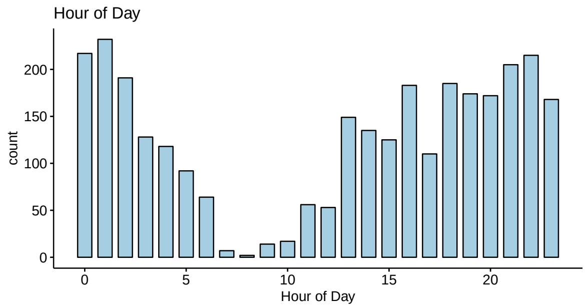
Censorship of Conservatives: @Google and @Twitter shadow-banning content and censoring ads ||

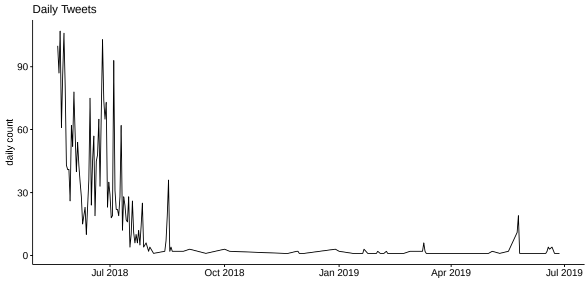
By RETWEETING this, is the only way to tell @sundarpichai and @jack they can't control and stop #FreeSpeech in #America

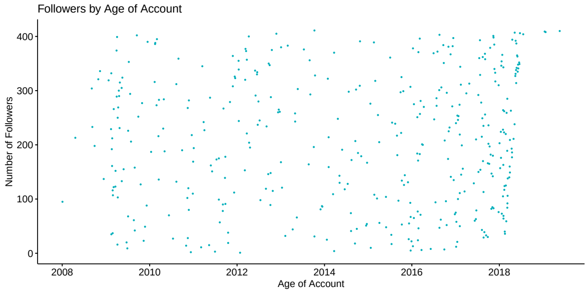
. **@TomTuTone1.** This account shows the confusing nature of the modern environment. It's description states "High profile bot. Available for disinformation campaigns, kids parties, bar and bat mitzvahs and football tailgate parties. Make Russia Dark Again". Probably not made by the Russians...or was it???









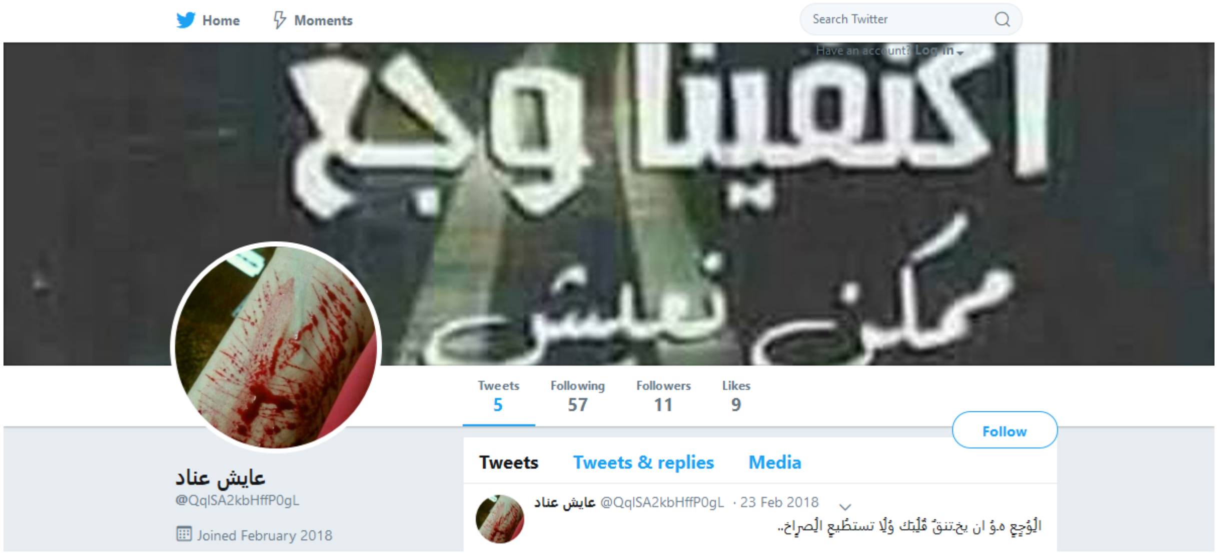


Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	1.523	bot-hunter Tier 0	False	H2P	24	realDonaldTrump	341	foxnews.com	10
Avg Sentiment (0-1)	0.598	bot-hunter Tier 1	0.586	TurnOnTheL19HTS	12	RomesburgJeremy	92	washex.am	9
Dormant Followers	13.1%	bot-hunter Tier 2	0.6048	Avenatti	10	Barnes_Law	89	fxn.ws	9
Bot Followers	21.2%	bot-hunter Tier 3	Deleted	Mueller	9	RKJ65	80	thehill.com	7
Avg Mentions	1.9	botometer	0.0608	MAGA	8	JuddLegum	77	ow.ly	6
Retweets RU Prop.	False	Debot	False	Pitt	7	AllisonRFloyd	67	dailycaller.com	5

E.8 Intimidation Bots

Intimidation bots are not designed to get friends. They are designed to follow designated target accounts in an attempt to push them off of Twitter. *Intimidation* accounts typically surge on the target account, following in large masses. These accounts can have disturbing images as their profile image.

In the next couple of pages we'll give some more tame examples of intimidation bots that we found attacking a journalist in the Middle East.

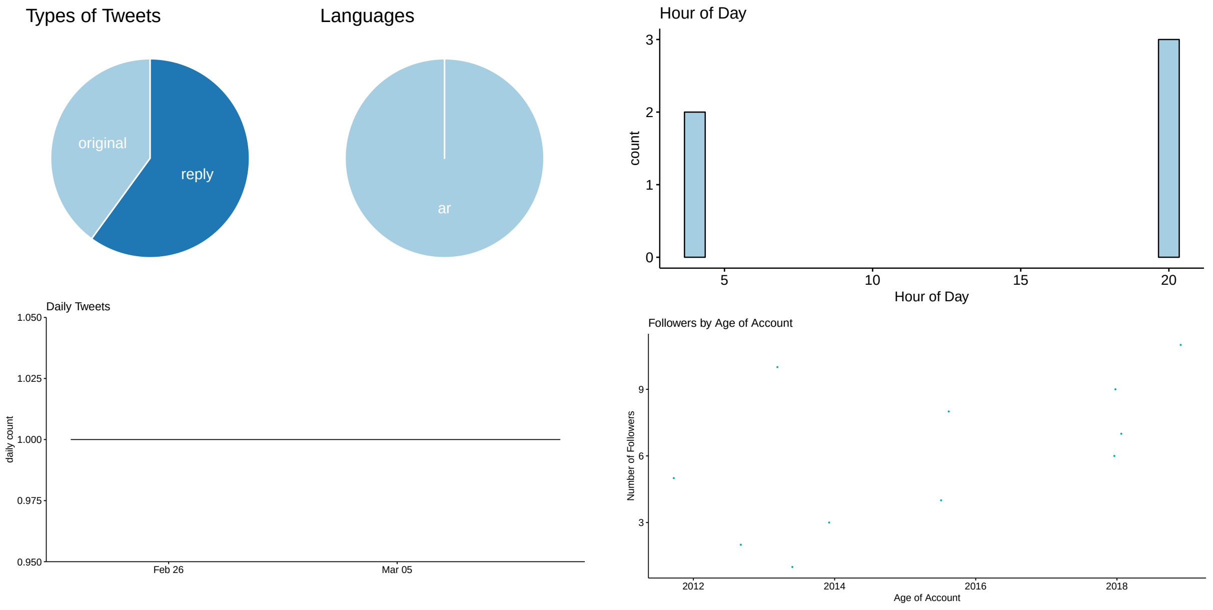


عايش عناد
@QqlSA2kbHffP0gL
Joined February 2018

عايش عناد @QqlSA2kbHffP0gL · 23 Feb 2018

الْوَجِع هوَ ان يخـتنقْ قلْبك وْلَا تستطِيع الْصراخ..

. @QqlSA2kbHffP0gL. We found @QqlSA2kbHffP0gL conducting intimidation operations against a free lance journalist in Yemen. The “pinned” tweet for this account shows a video of a sniper attacking a woman, implying that similar harm may await those who do not cease and desist on Twitter. The translated text of the first pinned tweet with Arabic quote "The pain is to suffocate your heart and you can’t scream.." The account doesn’t tweet, it just follows individuals to intimidate them.

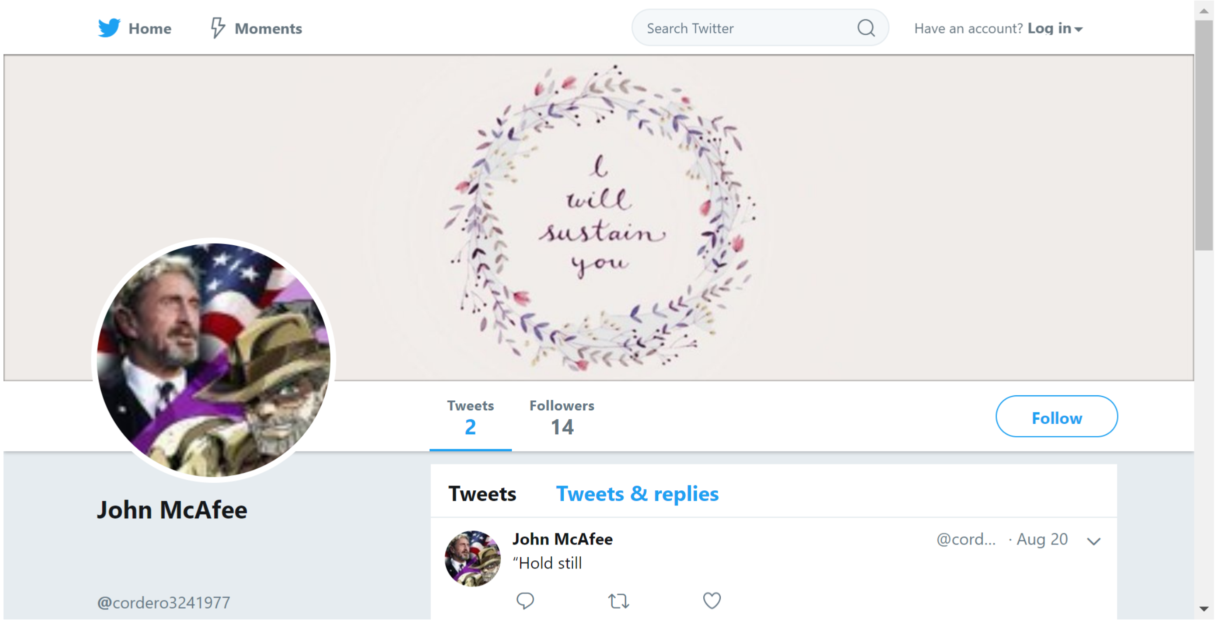


Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Count**	**Domain**
Frd/Fol Ratio	5.182	bot-hunter Tier 0	True	*feeling*	1.0	RandaAh82597393	1.0		
Avg Sentiment (0-1)	0.619	bot-hunter Tier 1	0.8264						
Dormant Followers	27.3%	bot-hunter Tier 2	0.6997						
Bot Followers	18.2%	bot-hunter Tier 3	0.299						
Avg Mentions	1	botometer	0.8762						
Retweets Russia Propaganda	False	Debot	False						

E.9 Coordinated Scam Accounts

Scam bots are accounts that work together to make a scam appear legitimate.

The next pages will present some examples of Bitcoin scam bots.



Home Moments

Search Twitter

Have an account? Log in

I will sustain you

Tweets 2

Followers 14

Follow

John McAfee

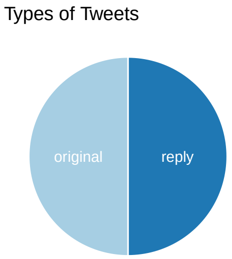
@cordero3241977

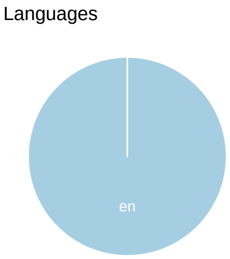
Tweets Tweets & replies

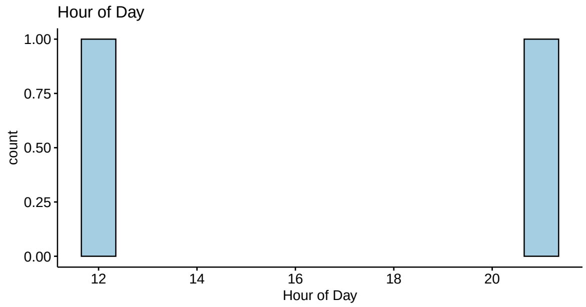
John McAfee @cord... · Aug 20

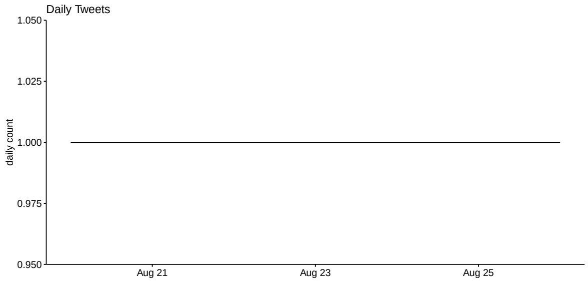
"Hold still

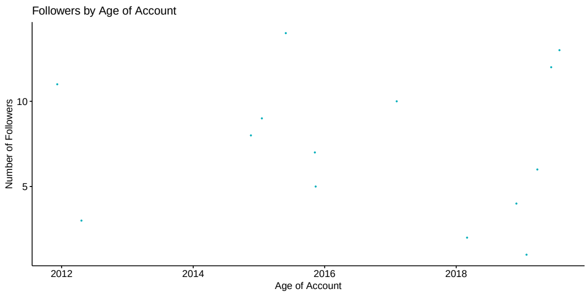
. **@cordero3241977.** We found @cordero3241977 linked to a Bitcoin bot scam often called the McAffee Bot Scam. This scam was able to steal 10's of thousands of dollars.



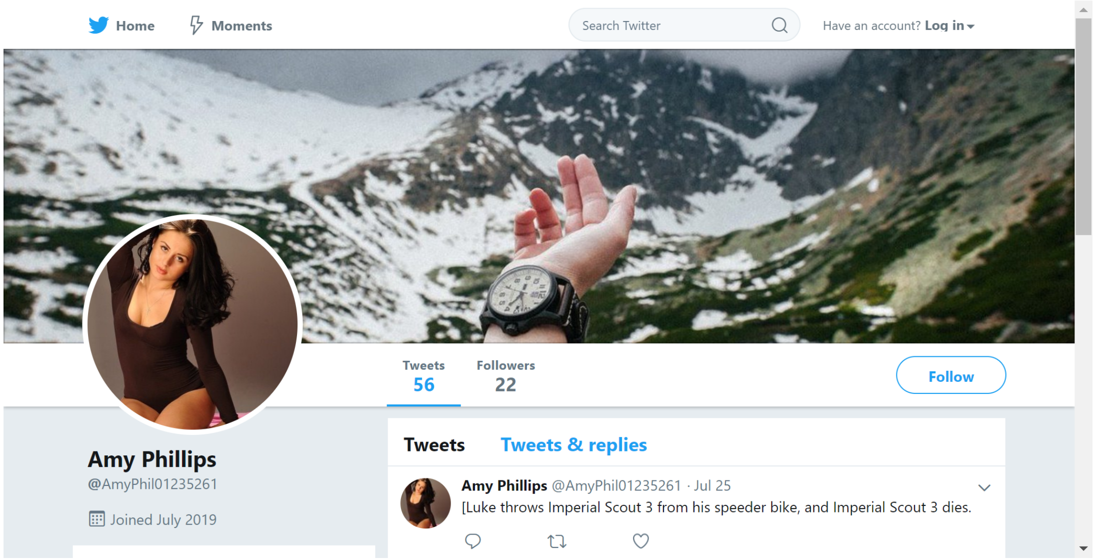








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False			realDonaldTrump	1.0		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.5577						
Dormant Followers	35.7%	bot-hunter Tier 2	0.6035						
Bot Followers	71.4%	bot-hunter Tier 3	Suspended						
Avg Mentions	1	botometer	0.7929						
Retweets RU Prop.	False	Debot	True						

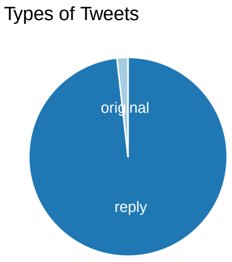


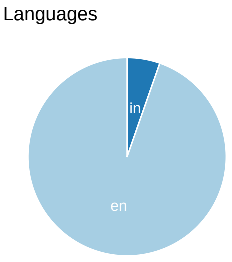
Amy Phillips
@AmyPhil01235261
Joined July 2019

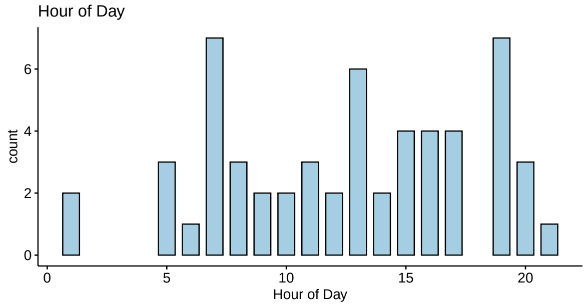
Tweets Tweets & replies

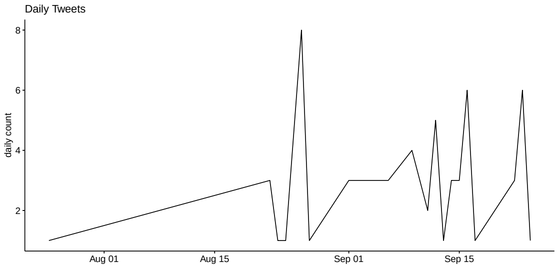
Amy Phillips @AmyPhil01235261 · Jul 25
[Luke throws Imperial Scout 3 from his speeder bike, and Imperial Scout 3 dies.

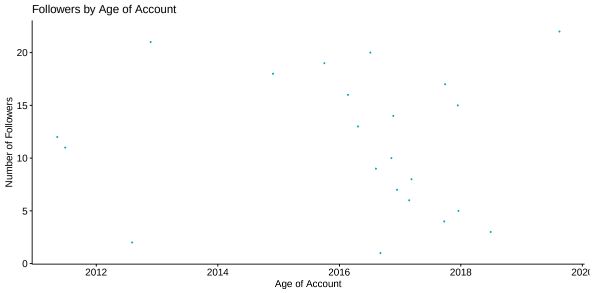
. @AmyPhil01235261. We found @AmyPhil01235261 linked to a McAffee Bot Scam that stole 10's of thousands of dollars. The @AmyPhil01235261 account would reply to the scam attempting to give it legitimacy. Tweets were "thank you Bakkt! got my 7 BTC" and "Great stuff! Gonna tell my friends asap." As seen here, many of these were *honey-pot* accounts.



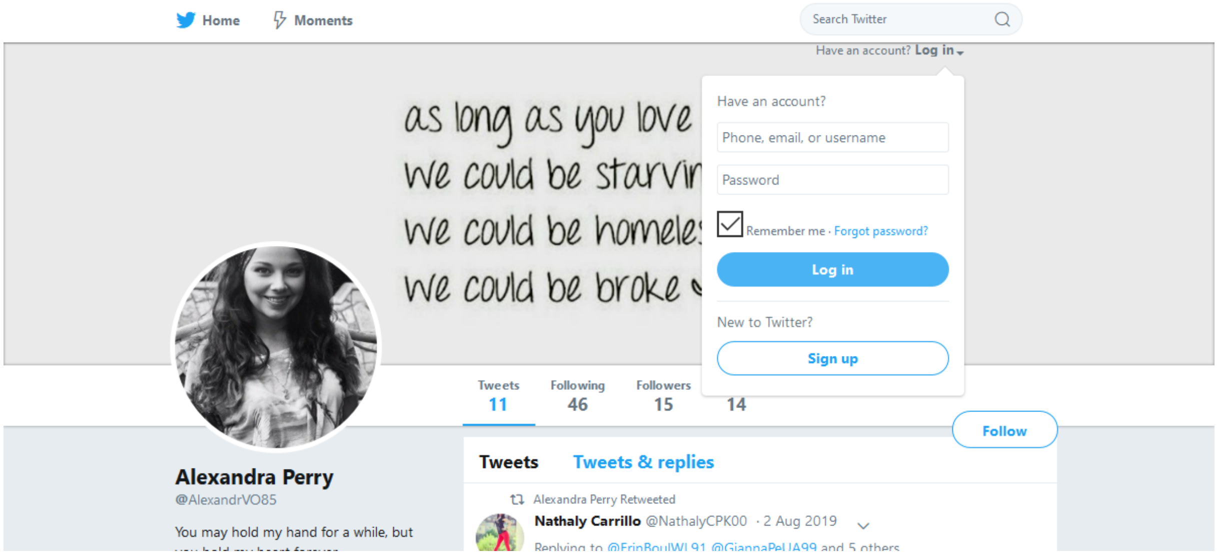








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0	bot-hunter Tier 0	False			realDonaldTrump	16		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.308			elonmusk	15		
Dormant Followers	36.4%	bot-hunter Tier 2	0.3088			Bakkt	14		
Bot Followers	77.3%	bot-hunter Tier 3	Suspended			Patrici27351679	5		
Avg Mentions	3.345	botometer	0.5968			MadeleinTN90	5		
Retweets RU Prop.	False	Debot	True			MokeSahri	4		

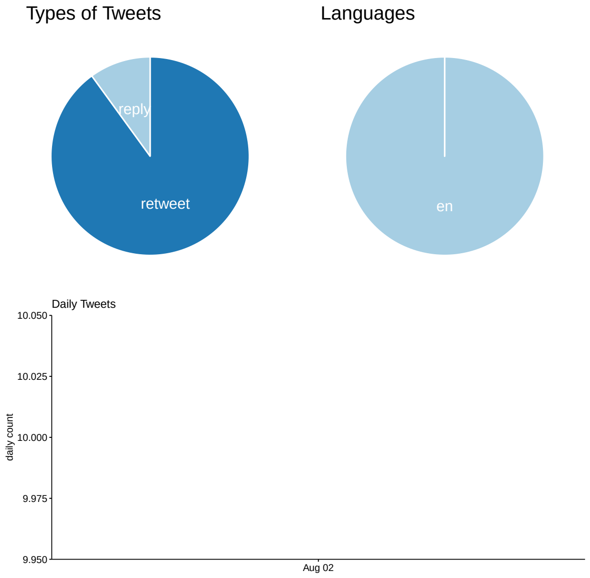


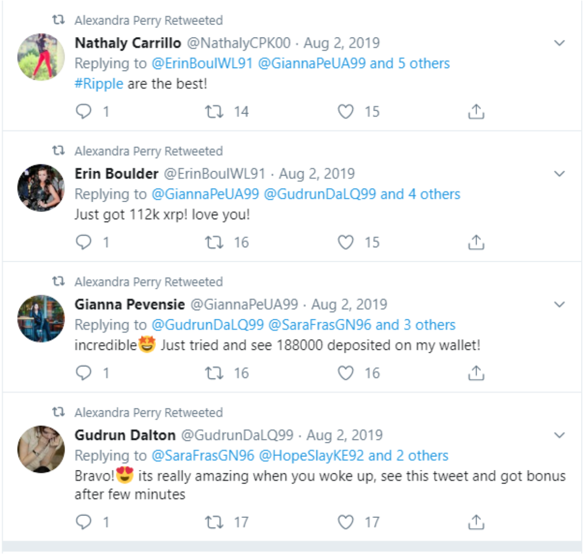
Alexandra Perry
@AlexandrVO85

You may hold my hand for a while, but

Alexandra Perry Retweeted
Nathaly Carrillo @NathalyCPK00 · 2 Aug 2019
Replying to @ErinBoulWL91 @GiannaPeUA99 and 5 others

. **@AlexandrVO85.** We found @AlexandrVO85 linked to the McAfee Bitcoin Scam.





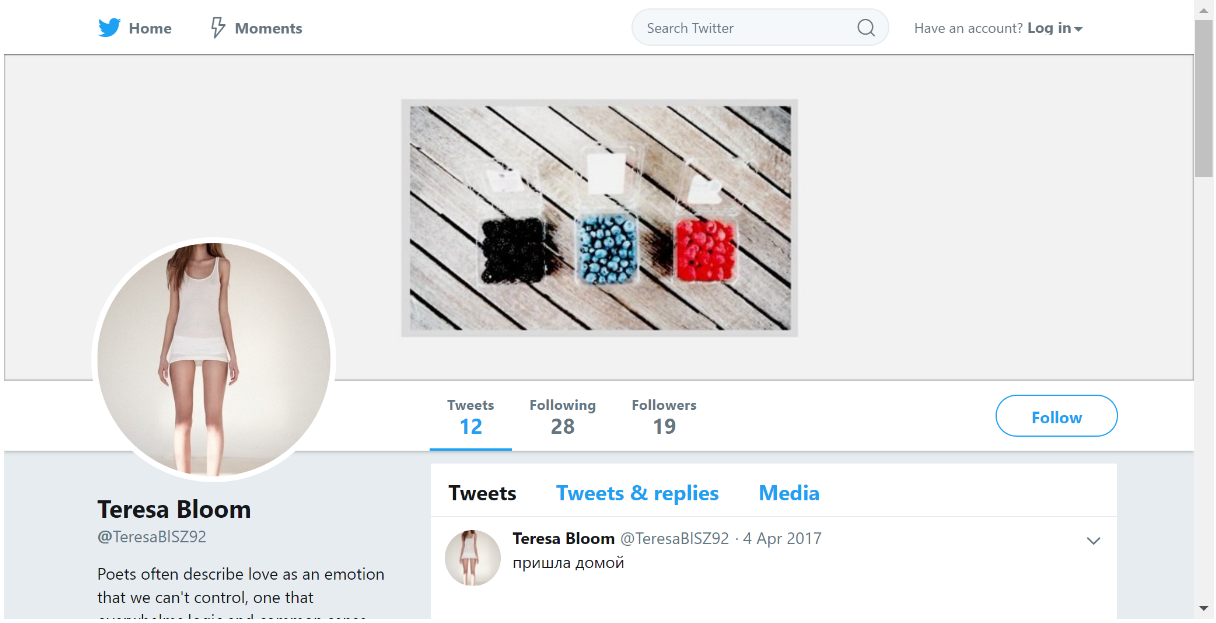
Alexandra Perry Retweeted
Nathaly Carrillo @NathalyCPK00 · Aug 2, 2019
Replying to @ErinBoulWL91 @GiannaPeUA99 and 5 others
#Ripple are the best!
1 14 15

Alexandra Perry Retweeted
Erin Boulder @ErinBoulWL91 · Aug 2, 2019
Replying to @GiannaPeUA99 @GudrunDaLQ99 and 4 others
Just got 112k xrp! love you!
1 16 15

Alexandra Perry Retweeted
Gianna Pevensie @GiannaPeUA99 · Aug 2, 2019
Replying to @GudrunDaLQ99 @SaraFrasGN96 and 3 others
incredible😍 Just tried and see 188000 deposited on my wallet!
1 16 16

Alexandra Perry Retweeted
Gudrun Dalton @GudrunDaLQ99 · Aug 2, 2019
Replying to @SaraFrasGN96 @HopeSlayKE92 and 2 others
Bravo!😍 its really amazing when you woke up, see this tweet and got bonus after few minutes
1 17 17

Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	3.067	bot-hunter Tier 0	False	*Ripple*	1.0	SaraFrasGN96	8		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.692			HopeSlayKE92	8		
Dormant Followers	33.3%	bot-hunter Tier 2	0.4561			GudrunDaLQ99	7		
Bot Followers	73.3%	bot-hunter Tier 3	0.541			michellebond111	7		
Avg Mentions	7.4	botometer	0.7990			Ripple	6		
Retweets RU Prop.	False	Debot	False			GiannaPeUA99	6		

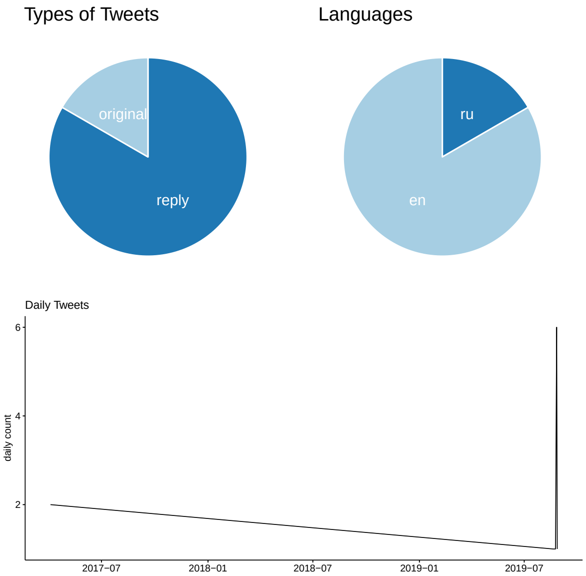


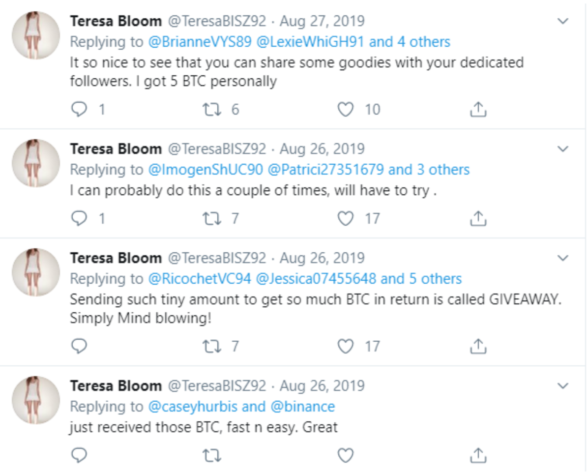
Teresa Bloom
@TeresaBlSZ92

Poets often describe love as an emotion that we can't control, one that

Teresa Bloom @TeresaBlSZ92 · 4 Apr 2017
пришла домой

. **@TeresaBISZ92.** We found @TeresaBlSZ92 linked to the McAfee Bitcoin Scam.





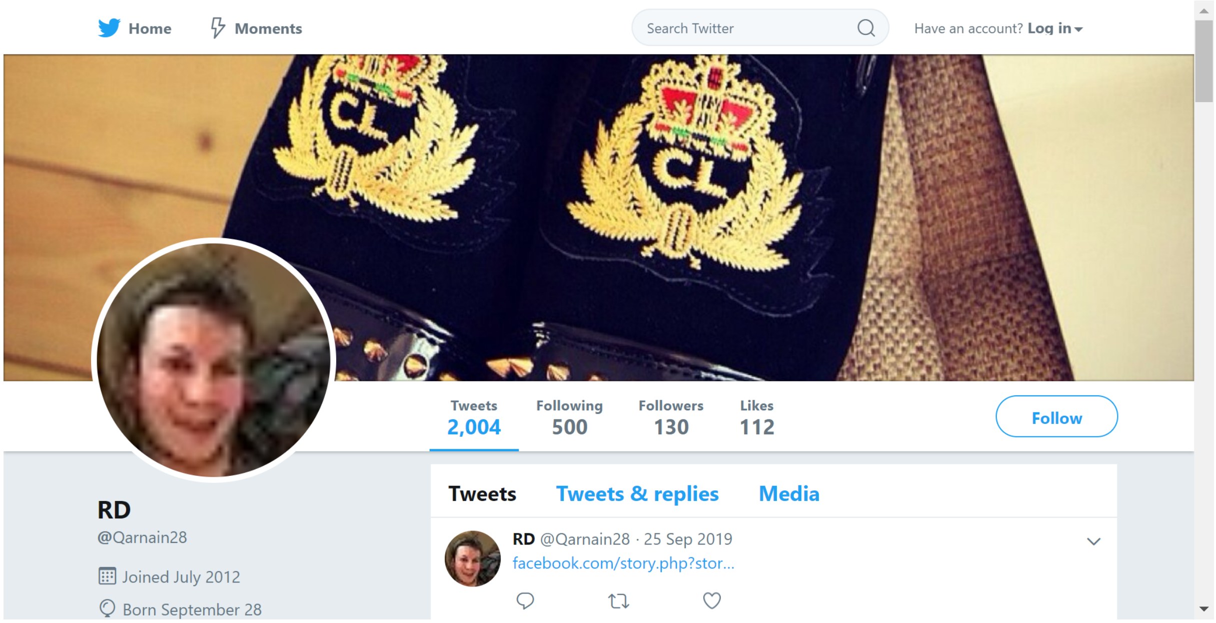
Teresa Bloom @TeresaBlSZ92 · Aug 27, 2019
Replying to @BrianneVYS89 @LexieWhiGH91 and 4 others
It so nice to see that you can share some goodies with your dedicated followers. I got 5 BTC personally
1 6 10

Teresa Bloom @TeresaBlSZ92 · Aug 26, 2019
Replying to @ImogenShUC90 @Patrici27351679 and 3 others
I can probably do this a couple of times, will have to try .
1 7 17

Teresa Bloom @TeresaBlSZ92 · Aug 26, 2019
Replying to @RicochetVC94 @Jessica07455648 and 5 others
Sending such tiny amount to get so much BTC in return is called GIVEAWAY. Simply Mind blowing!
7 17

Teresa Bloom @TeresaBlSZ92 · Aug 26, 2019
Replying to @caseyhurbis and @binance
just received those BTC, fast n easy. Great

Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	1.474	bot-hunter Tier 0	False			realDonaldTrump	7	pscp.tv	1.0
Avg Sentiment (0-1)		bot-hunter Tier 1	0.5781			LexieWhiGH91	5		
Dormant Followers	31.6%	bot-hunter Tier 2	0.4199			RicochetVC94	4		
Bot Followers	57.9%	bot-hunter Tier 3	Suspended			BrianneVYS89	3		
Avg Mentions	5.2	botometer	0.7713			ImogenShUC90	3		
Retweets RU Prop.	False	Debot	False			Jessica07455648	2		



Home Moments Search Twitter Have an account? Log in

Tweets 2,004 Following 500 Followers 130 Likes 112 Follow

Tweets Tweets & replies Media

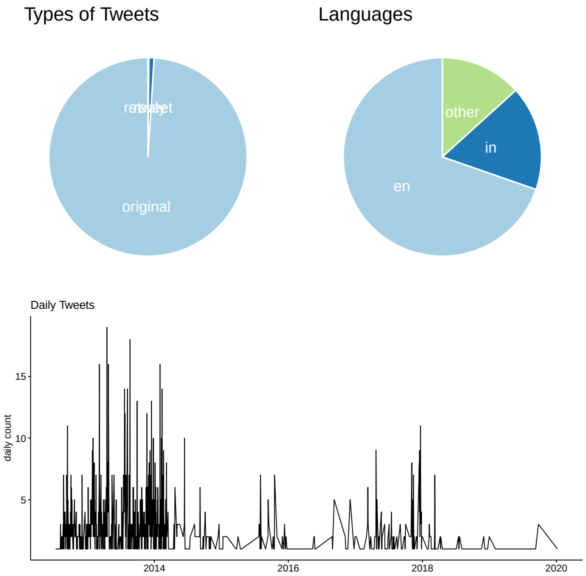
RD
@Qarnain28

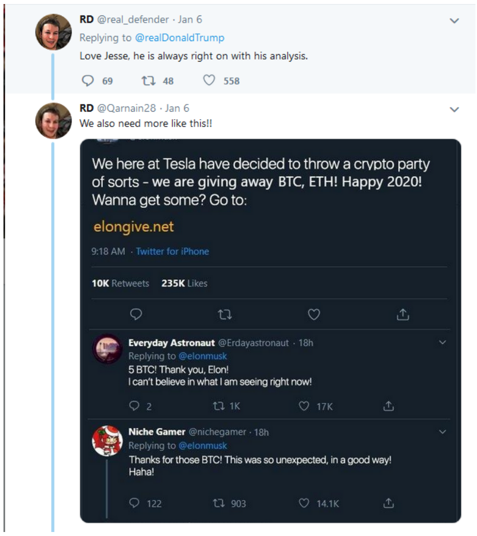
Joined July 2012

Born September 28

RD @Qarnain28 · 25 Sep 2019
facebook.com/story.php?stor...

. **@Qarnain28.** We found @Qarnain28 linked to the McAfee Bitcoin Scam.





RD @real_defender · Jan 6
Replying to @realDonaldTrump
Love Jesse, he is always right on with his analysis.
69 48 558

RD @Qarnain28 · Jan 6
We also need more like this!!

We here at Tesla have decided to throw a crypto party of sorts - we are giving away BTC, ETH! Happy 2020! Wanna get some? Go to:

elongive.net

9:18 AM · Twitter for iPhone

10K Retweets 235K Likes

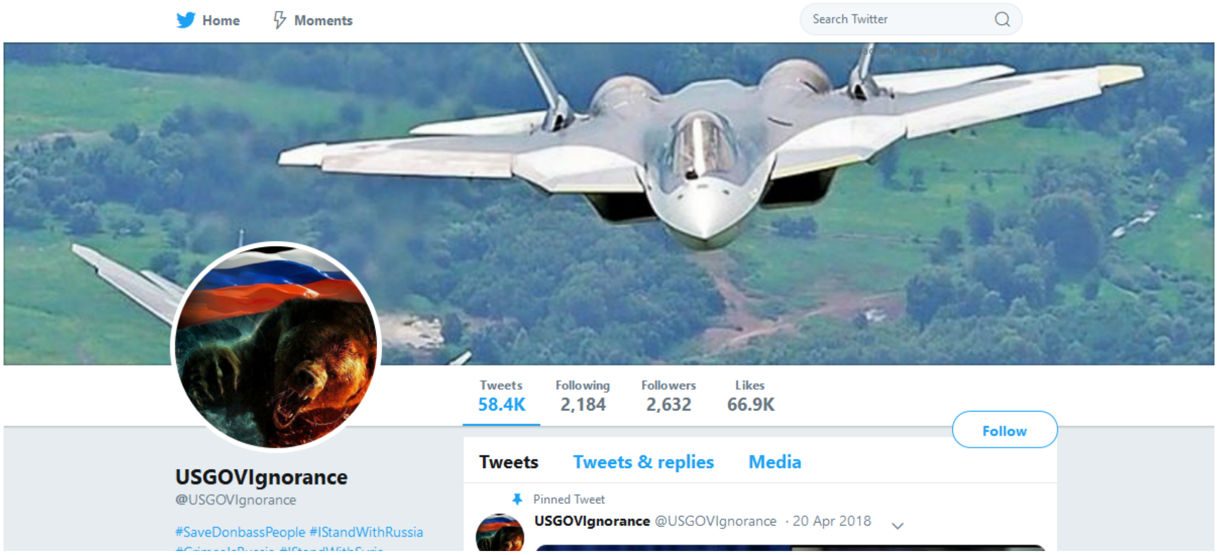
Everyday Astronaut @Erdayastronaut · 18h
Replying to @elonmusk
5 BTC! Thank you, Elon!
I can't believe in what I am seeing right now!
2 1K 17K

Niche Gamer @nichegamer · 18h
Replying to @elonmusk
Thanks for those BTC! This was so unexpected, in a good way! Haha!
122 903 14.1K

Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	1.474	bot-hunter Tier 0	False	*bruneimua*	70	mohammad$_{a}sy$	9	fb.me	673
Avg Sentiment (0-1)		bot-hunter Tier 1	0.407	*bruneimakeupartist*	64	mdasy1	8	instagram.com	664
Dormant Followers	31.6%	bot-hunter Tier 2	0.3131	*muabrunei*	61	BebyG	8	instagr.am	450
Bot Followers	57.9%	bot-hunter Tier 3	0.220	*weddingbrunei*	55	jz355	7	invite4job.com	12
Avg Mentions	1.211	botometer	0.5566	*getmerated*	54	AyaiWinchester	7	facebook.com	7
Retweets RU Prop.	False	Debot	False	*instacollage*	49	CeliaRahim	6	starmakerstudios.com	2

E.10 Russian, Iranian, and Middle East Bots

The following pages go into bots, trolls, and cyborgs that we've found associated with Russia, Iran, China or the conflict in Yemen.

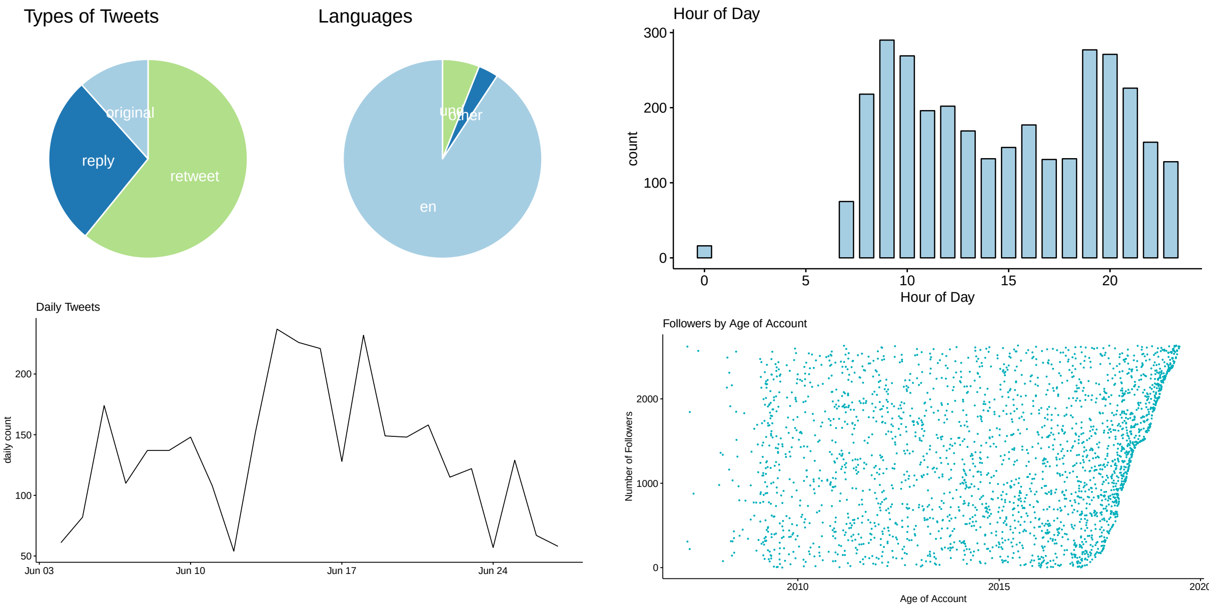


USGOVIgnorance
@USGOVIgnorance

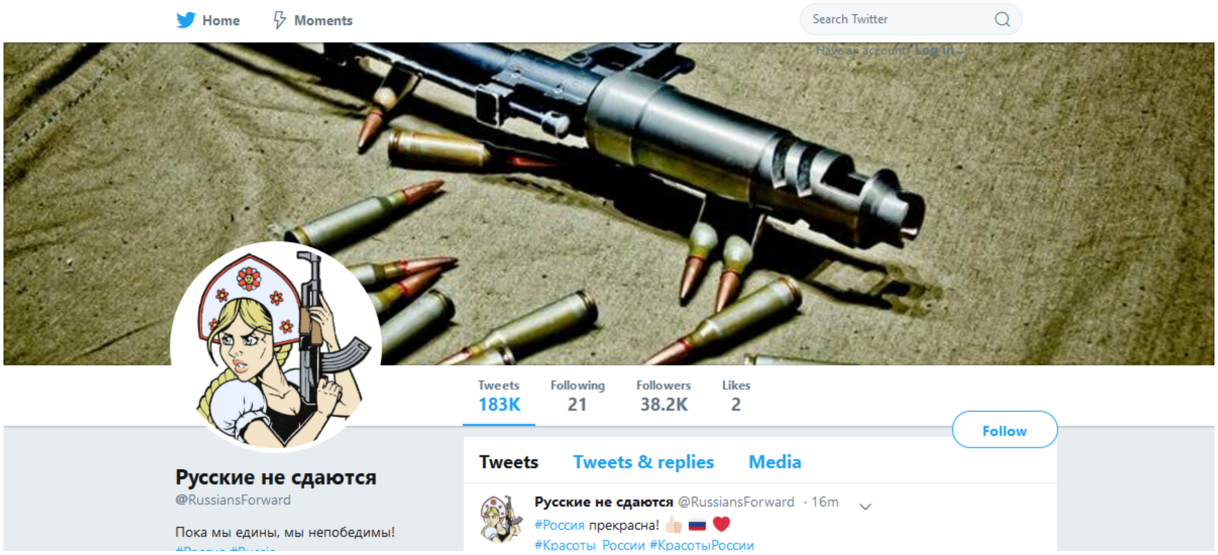
#SaveDonbassPeople #IStandWithRussia

USGOVIgnorance @USGOVIgnorance · 20 Apr 2018

. **@USGOVIgnorance.** Overt Russian Propaganda Account (or an account masquerading as a Russian propaganda account)



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.83	bot-hunter Tier 0	False	US	69	9arsth	281	rt.com	36
Avg Sentiment (0-1)	0.565	bot-hunter Tier 1	0.8421	Ukraine	59	Immort4l$_{Legacy}$	227	sputniknews.com	23
Dormant Followers	14.4%	bot-hunter Tier 2	0.7556	Iran	56	KermitHigby	219	youtu.be	15
Bot Followers	27.7%	bot-hunter Tier 3	0.551	Russia	51	BoomerangTime	200	youtube.com	12
Avg Mentions	2.761	botometer	0.0959	Donetsk	27	MisterTwyst	199	bit.ly	10
Retweets RU Prop.	True	Debot	False	DPR	25	USGOVIgnorance	186	tass.com	8

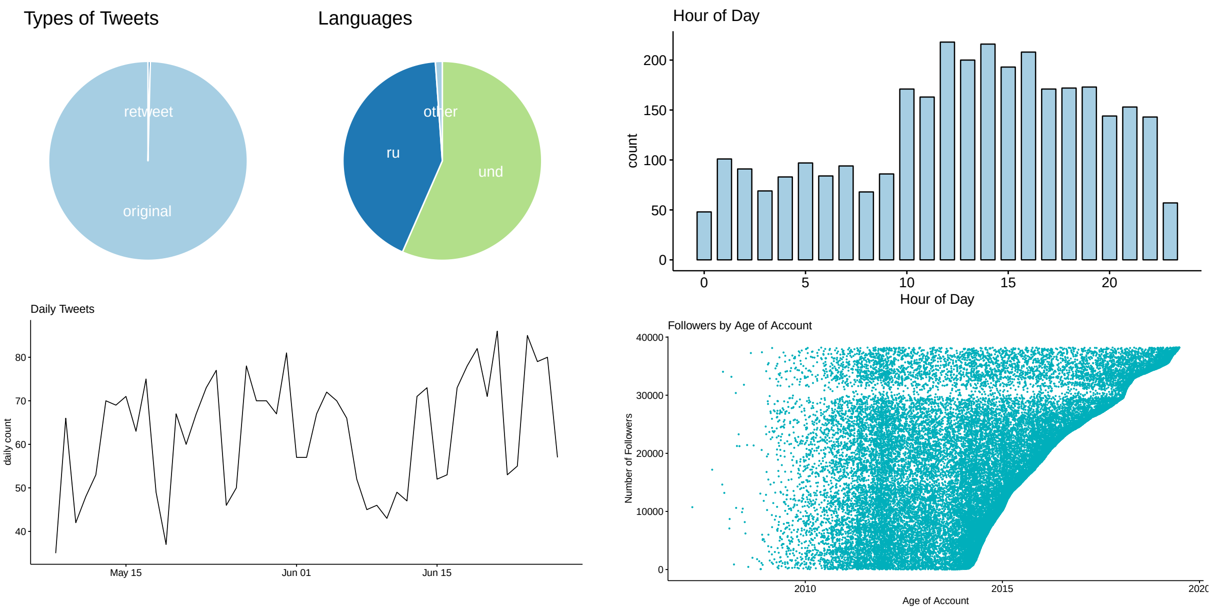


Русские не сдаются
@RussiansForward

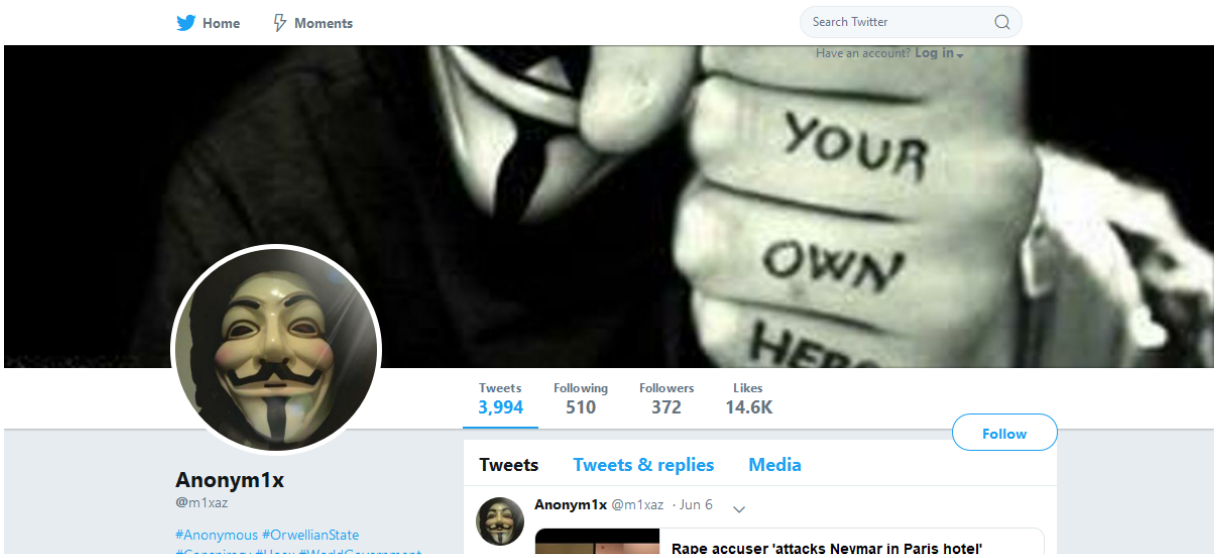
Пока мы едины, мы непобедимы!

Русские не сдаются @RussiansForward · 16m
#Россия прекрасна! 👍 🇷🇺 ❤
#Красоты_России #КрасотыРоссии

. **@RussiansForward.** This is an overt Russian propaganda account



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.001	bot-hunter Tier 0	False	*Humor*	765	good$_{e}vents$	7.0	youtu.be	130
Avg Sentiment (0-1)	0.639	bot-hunter Tier 1	0.1479	*History*	505	Current_policy	3.0	youtube.com	66
Dormant Followers	52.5%	bot-hunter Tier 2	0.4493	*Our heroes*	244	russia_sila	1.0	vk.com	34
Bot Followers	50.4%	bot-hunter Tier 3	0.409	*Defense*	110	20let$_{n}azad$	1.0	tinyurl.com	12
Avg Mentions	1	botometer	0.138	*Awareness*	85			ria.ru	5
Retweets Russia Propaganda	False	Debot	False	*Russia*	71			vk.cc	4



Tweets 3,994 | Following 510 | Followers 372 | Likes 14.6K

Anonym1x
@m1xaz

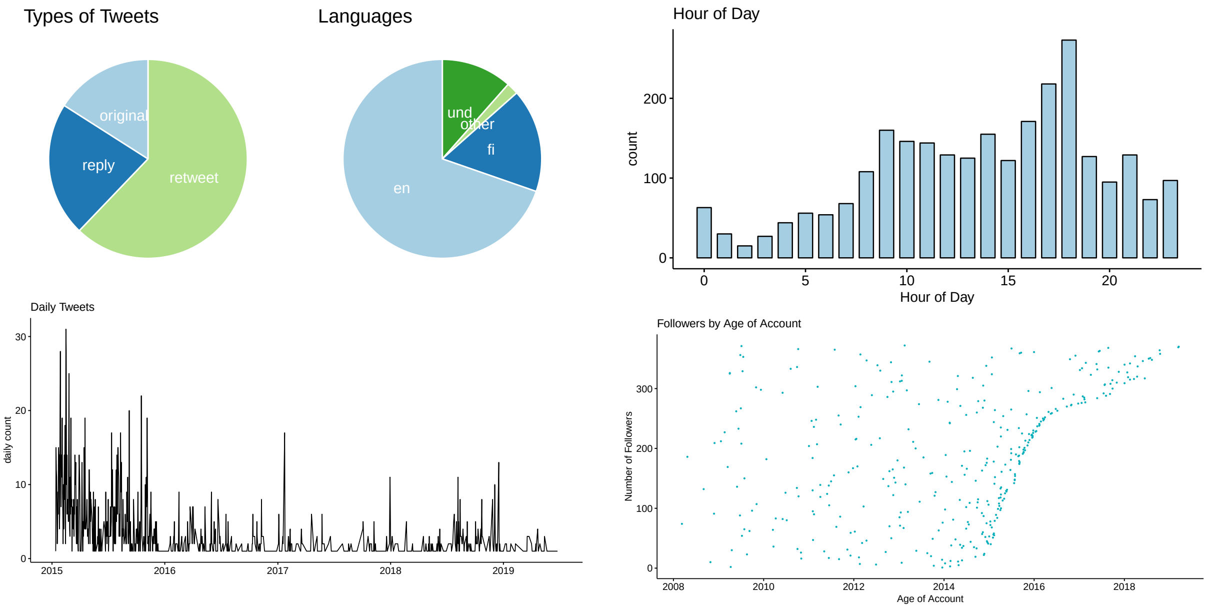
#Anonymous #OrwellianState

Tweets | Tweets & replies | Media

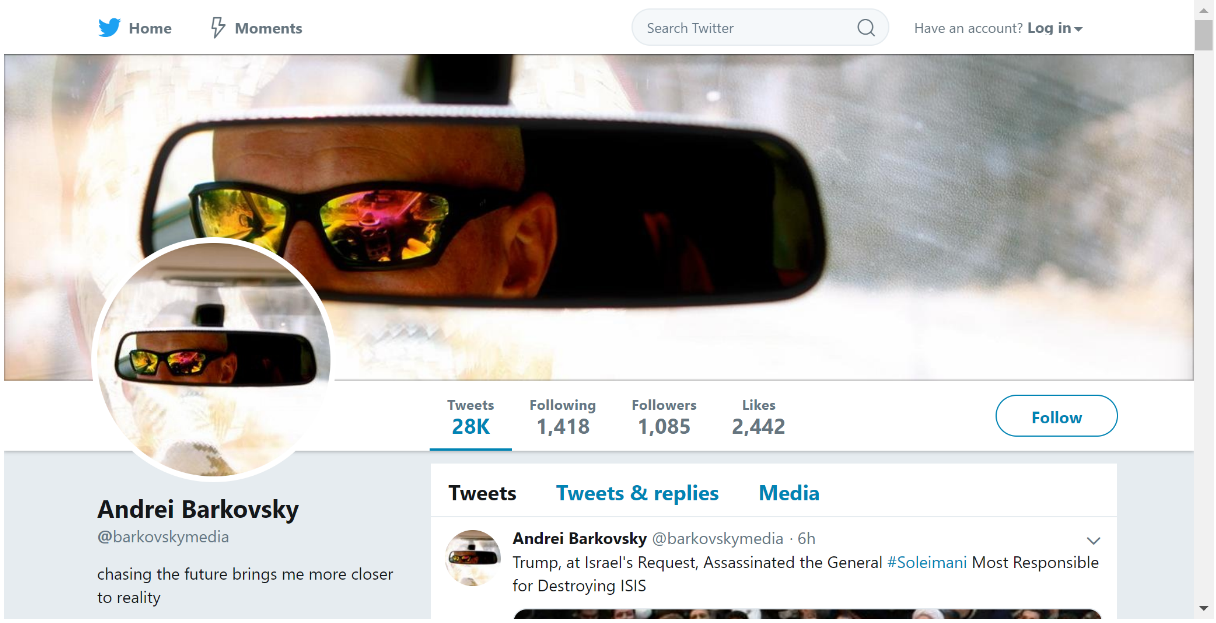
Anonym1x @m1xaz · Jun 6

Rape accuser 'attacks Neymar in Paris hotel'

. **@m1xaz.** Account has ties to Anonymous messaging and possibly Russian messaging



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	1.371	bot-hunter Tier 0	False	Anonymous	53	davidicke	450	bit.ly	306
Avg Sentiment (0-1)	0.592	bot-hunter Tier 1	0.308	OpDeathEaters	33	VERKKOMEDIAorg	118	rt.com	102
Dormant Followers	44.1%	bot-hunter Tier 2	0.5757	Russia	27	RT_com	97	fb.me	96
Bot Followers	27.7%	bot-hunter Tier 3	0.627	mitävittua	27	MassDeception1	65	youtu.be	54
Avg Mentions	1.296	botometer	0.0445	ISIS	26	JuhaMunposti	63	sptnkne.ws	31
Retweets RU Prop.	True	Debot	False	US	25	mitavittualehti	56	b4in.org	31



Tweets 28K | Following 1,418 | Followers 1,085 | Likes 2,442

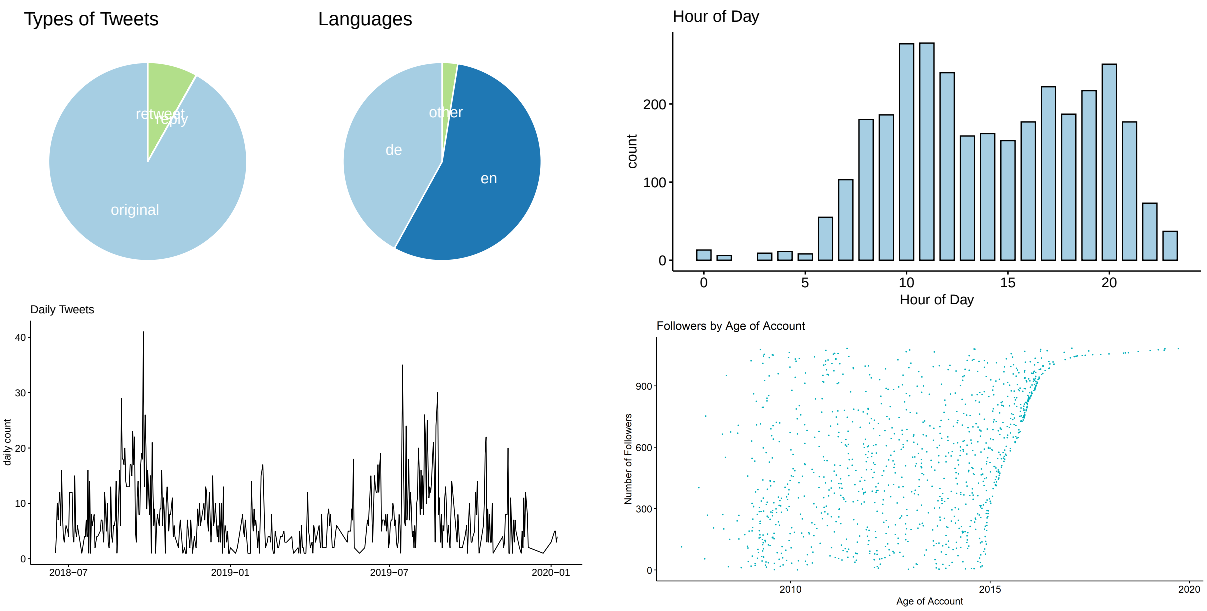
Tweets | Tweets & replies | Media

Andrei Barkovsky
@barkovskymedia

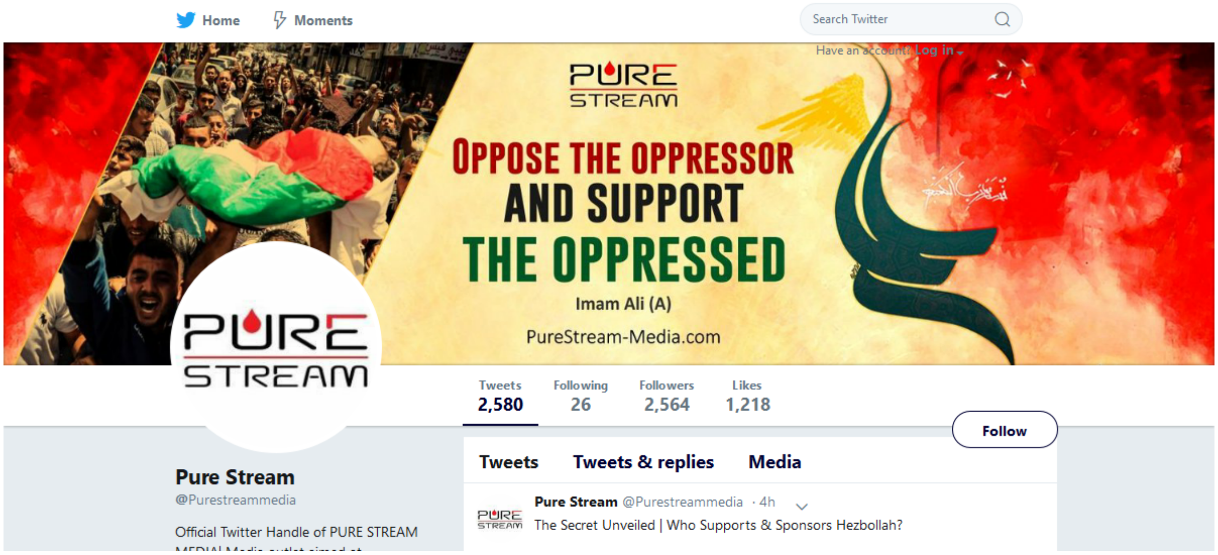
chasing the future brings me more closer to reality

Andrei Barkovsky @barkovskymedia · 6h
Trump, at Israel's Request, Assassinated the General #Soleimani Most Responsible for Destroying ISIS

. **@barkovskymedia.** We found @barkovskymedia primarily just amplifies Russian State Sponsored News Outlets.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	1.307	bot-hunter Tier 0	False	*US*	188	de$_{sputnik}$	691	sputniknews.com	1302
Avg Sentiment (0-1)		bot-hunter Tier 1	0.749	*Merkel*	144	SputnikInt	599	tagesschau.de	96
Dormant Followers	37.8%	bot-hunter Tier 2	0.5095	*I*	47	YouTube	22	rt.com	87
Bot Followers	26.5%	bot-hunter Tier 3	0.534	*Trump*	47	NOS	18	theduran.com	28
Avg Mentions	1.029	botometer	0.1701	*UK*	46	DWN$_{de}$	12	youtu.be	22
Retweets RU Prop.	True	Debot	False	*NATO*	29	mazzenilsson	8	nos.nl	19



Home Moments

Search Twitter

Have an account? Log in

PURE STREAM

OPPOSE THE OPPRESSOR AND SUPPORT THE OPPRESSED

Imam Ali (A)

PureStream-Media.com

PURE STREAM

Tweets	Following	Followers	Likes
2,580	26	2,564	1,218

Follow

Pure Stream

@Purestreammedia

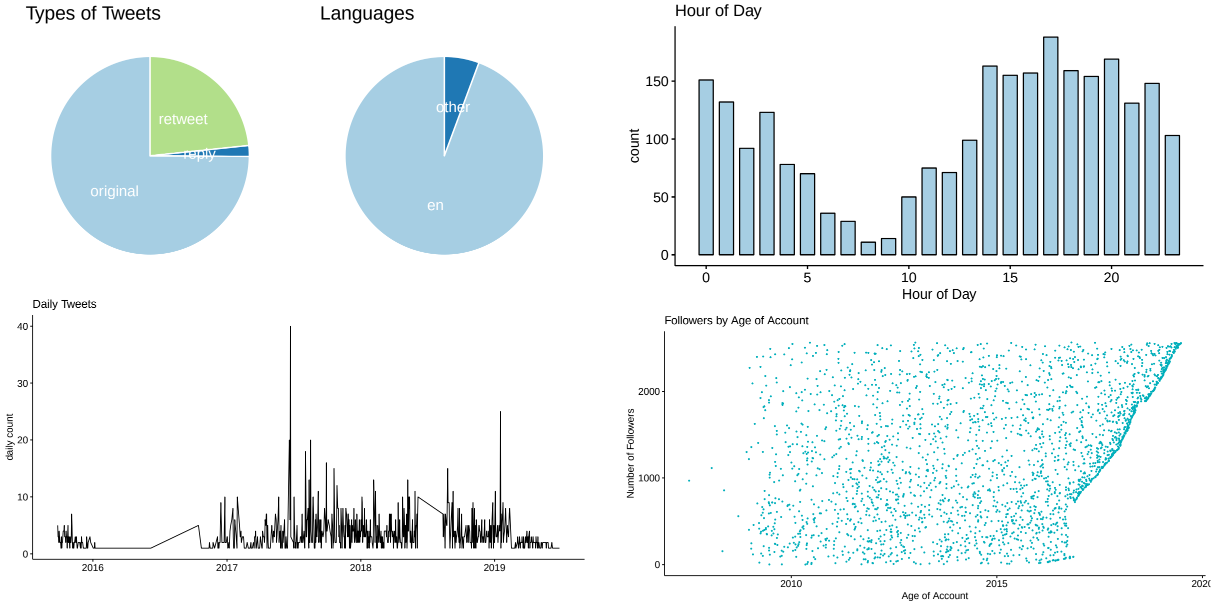
Official Twitter Handle of PURE STREAM

Tweets Tweets & replies Media

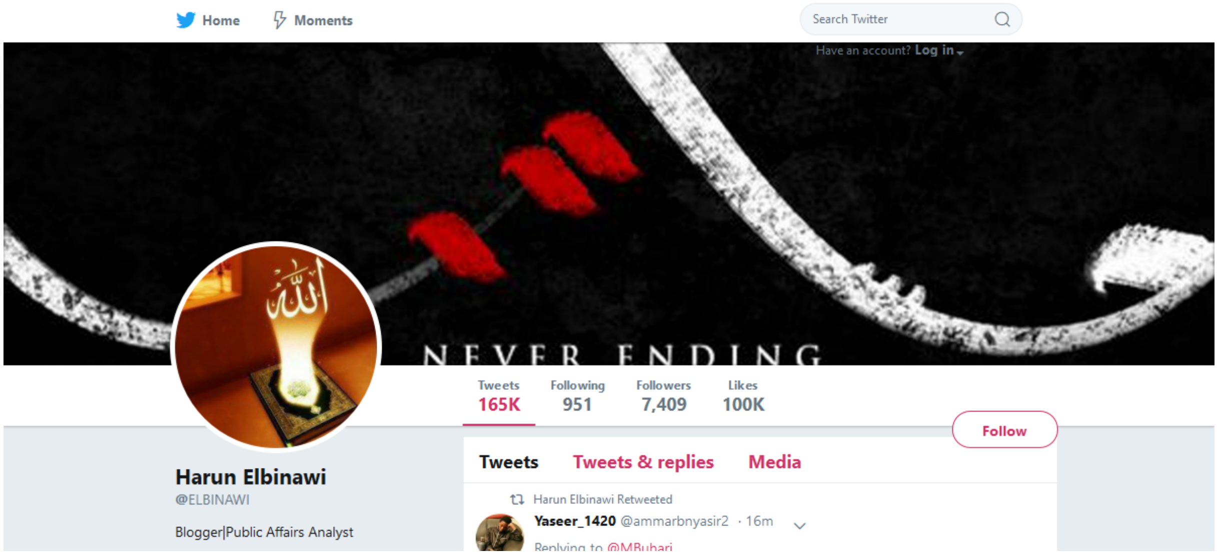
Pure Stream @Purestreammedia · 4h

The Secret Unveiled | Who Supports & Sponsors Hezbollah?

. **@Purestreammedia.** This is an overt Iranian propaganda account.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.01	bot-hunter Tier 0	False	America	103	Purestreammedia	437	fb.me	139
Avg Sentiment (0-1)	0.607	bot-hunter Tier 1	0.2915	Palestine	99	khamenei_ir	61	purestream-media.com	98
Dormant Followers	36.6%	bot-hunter Tier 2	0.3072	Islam	83	islamic_pulse	33	Instagram.com	47
Bot Followers	47.7%	bot-hunter Tier 3	0.438	ImamKhamenei	78	Qom_TV	19	fb.com	45
Avg Mentions	1.238	botometer	0.6729	Iran	59	TeamConfronters	15	Telegram.me	33
Retweets RU Prop.	False	Debot	False	DeathToAmerica	56	aimislam	13	Fb.com	22



Tweets 165K | Following 951 | Followers 7,409 | Likes 100K

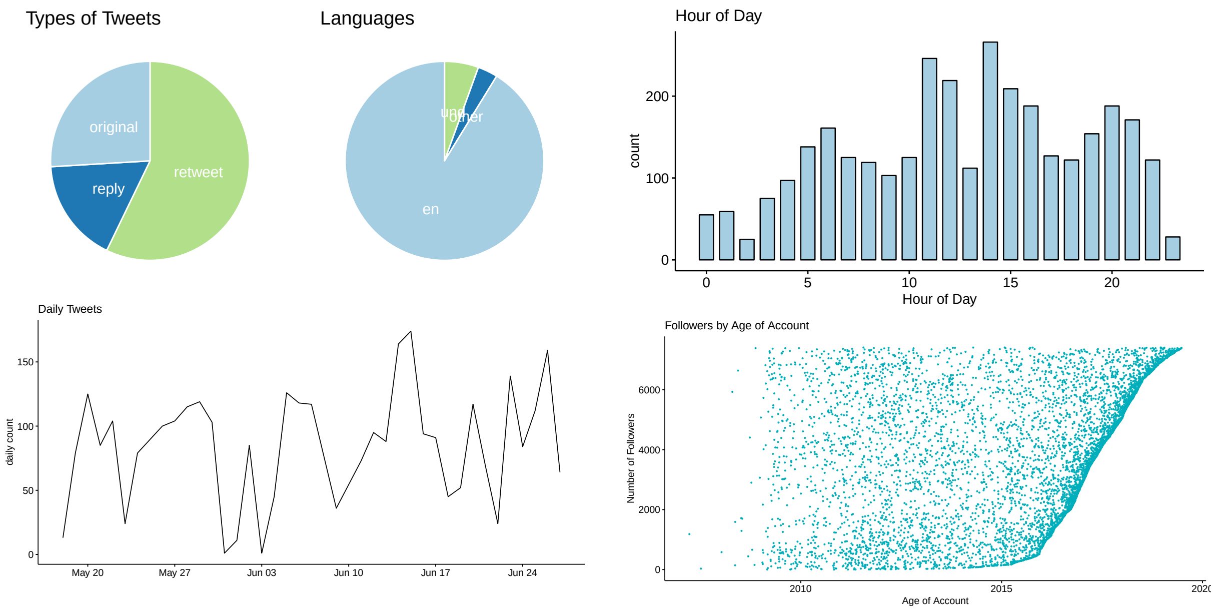
Tweets | Tweets & replies | Media

Harun Elbinawi
@ELBINAWI

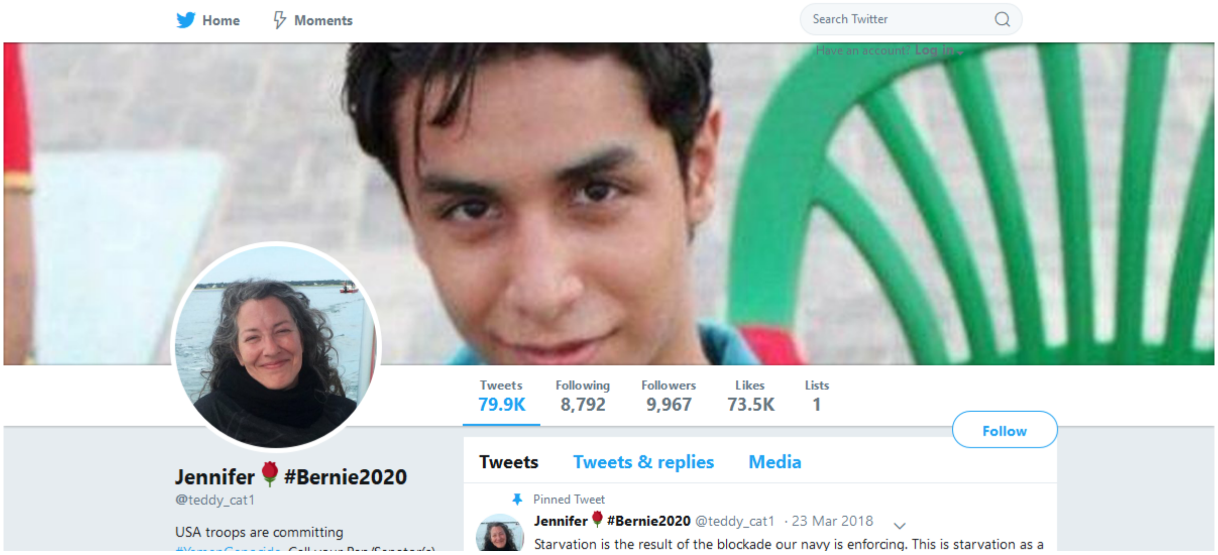
Blogger|Public Affairs Analyst

Harun Elbinawi Retweeted
Yaseer_1420 @ammarbnyasir2 · 16m
Replying to @MBuhari

. **@ELBINAWI.** This account states it is located in Nigeria, but is definitely a bot and has strong ties to Iran.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.128	bot-hunter Tier 0	False	FreeZakzaky	81	MBuhari	347	dailytrust.com.ng	19
Avg Sentiment (0-1)	0.551	bot-hunter Tier 1	0.5148	Iran	66	elrufai	99	wordpress.com	18
Dormant Followers	46.6%	bot-hunter Tier 2	0.7703	ZariaMassacre	66	yusufhsani	83	bit.ly	15
Bot Followers	45.1%	bot-hunter Tier 3	0.654	ZariaGenocide	53	AmnestyNigeria	75	almanar.com.lb	15
Avg Mentions	1.64	botometer	0.0445	FreeZakZaky	49	ELBINAWI	73	ptv.io	13
Retweets RU Prop.	True	Debot	False	FreePalestine	38	khamenei_ir	67	aje.io	13

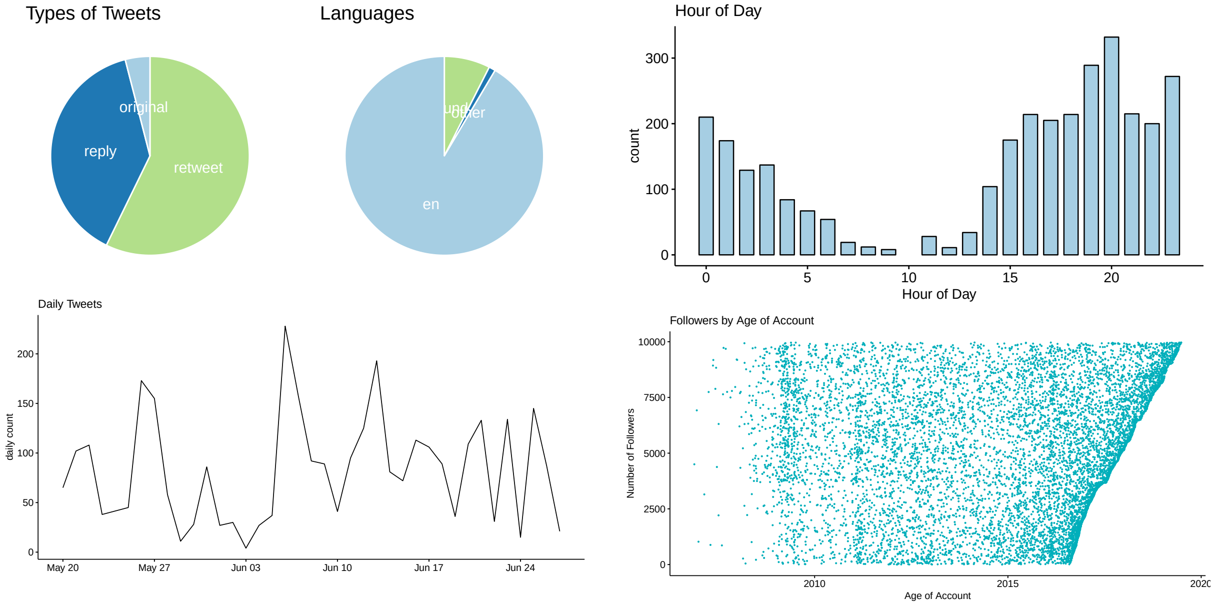


Jennifer 🌹 #Bernie2020
@teddy_cat1

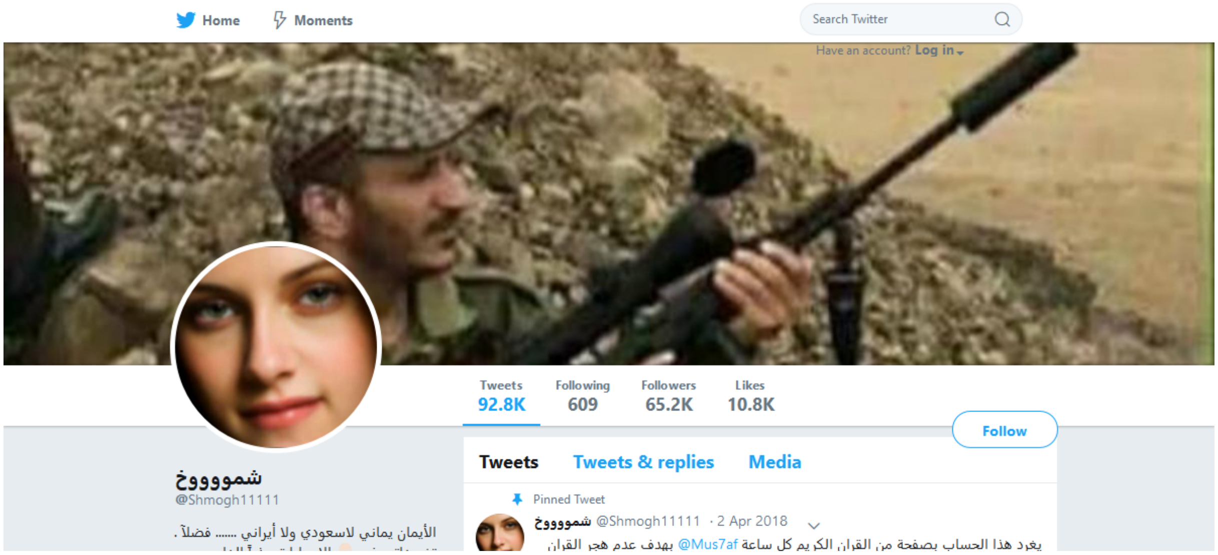
USA troops are committing
#YemenGenocide. Call your Rep/Senator(s)

Jennifer 🌹 #Bernie2020 @teddy_cat1 · 23 Mar 2018
Starvation is the result of the blockade our navy is enforcing. This is starvation as a

. **@teddy_cat1.** Has possible ties to Iran and/or Russia



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.882	bot-hunter Tier 0	False	Yemen	100	teddy_cat1	264	youtu.be	6
Avg Sentiment (0-1)	0.582	bot-hunter Tier 1	0.5694	Saudi	37	BernieSanders	236	theguardian.com	4
Dormant Followers	27.5%	bot-hunter Tier 2	0.9077	MedicareForAll	29	WalkerBragman	125	cnn.com	3
Bot Followers	34.8%	bot-hunter Tier 3	0.695	Bernie2020	28	40_Ronda	78	bit.ly	3
Avg Mentions	2.068	botometer	0.1481	YemenGenocide	28	DOJMainJustice	77	washingtonpost.com	3
Retweets RU Prop.	True	Debot	False	ClimateEmergency	17	XoXo__Kellie	76	rt.com	3



Home Moments Search Twitter

Have an account? Log in

Tweets	Following	Followers	Likes
92.8K	609	65.2K	10.8K

Follow

شموووووخ
@Shmogh11111

الأيمان يماني لاسعودي ولا أيراني .... فضلاً .

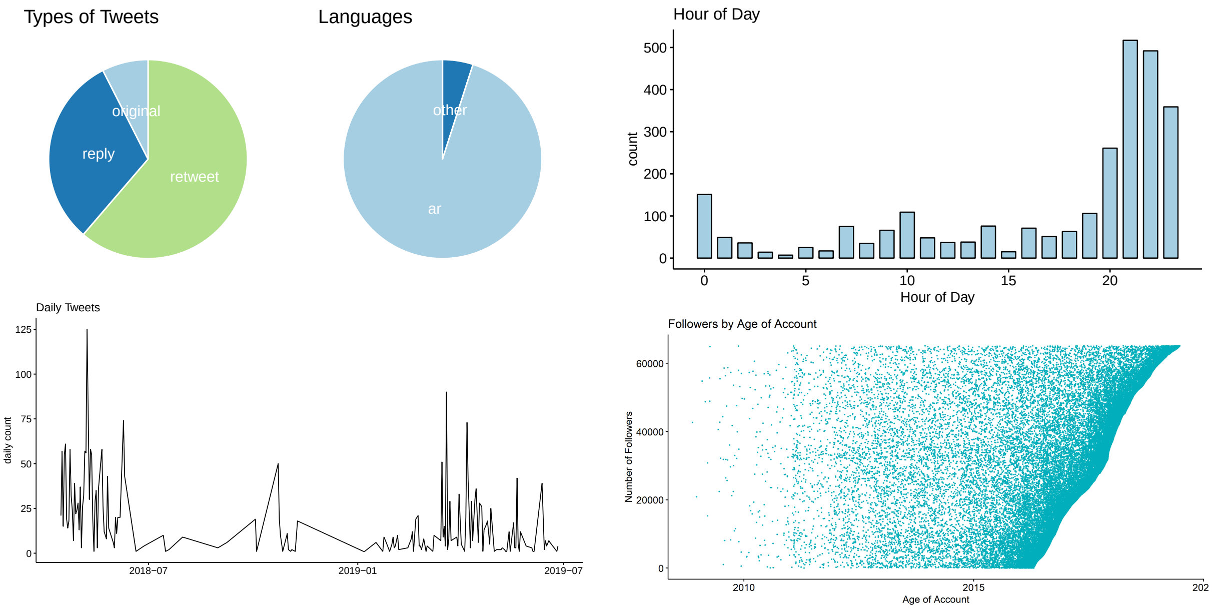
Tweets Tweets & replies Media

Pinned Tweet

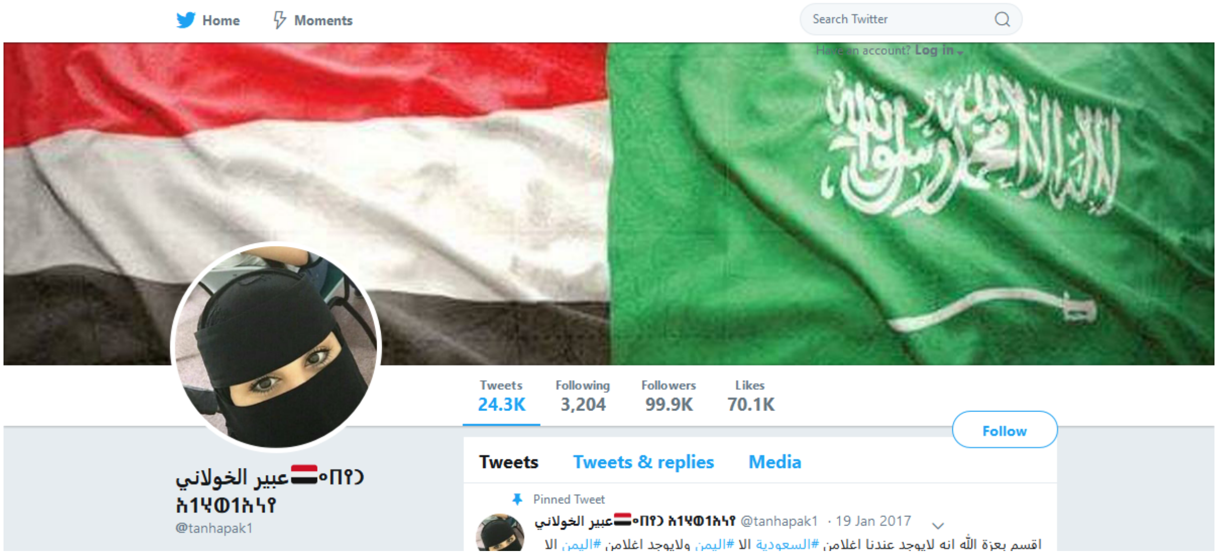
شموووووخ @Shmogh11111 · 2 Apr 2018

يغرد هذا الحساب بصفحة من القران الكريم كل ساعة @Mus7af بهدف عدم هجر القران

. @Shmogh11111. We found @Shmogh11111 associated with the Yemen conflict. The translated description on the account claims "This account is credited with a page of the Holy Quran every hour @Mus7af in order not to abandon the Holy Quran account worthy of follow-up." The account goes beyond the Quran and provides significant political commentary and news about Yemen.



Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Domain**	**C**
Frd/Fol Ratio	0.009	bot-hunter Tier 0	False	*To whom*	22	akramHajarr	116	2dec.net	
Avg Sentiment (0-1)	0.683	bot-hunter Tier 1	0.373	*Sana'a*	21	D9ytEX8aZHxHe9n	109	aja.me	
Dormant Followers	61.5%	bot-hunter Tier 2	0.595	*Guards of the republic*	21	67_5nsh$_{9}$5	70	facebook.com	
Bot Followers	45.8%	bot-hunter Tier 3	0.378	*Fatwa Hawthi kill our children*	20	heyam2255	58	youtu.be	
Avg Mentions	1.314	botometer	0.1947	*Houthi*	14	00Benyameen	58	rt.com	
Retweets Russia Propaganda	True	Debot	False	*Hodeidah*	13	Al_Affash	58	alathkar.org	



Home Moments

Search Twitter

account? Log in

Tweets 24.3K Following 3,204 Followers 99.9K Likes 70.1K

Follow

عبير الخولاني🇾🇪ⴰⵏⵉⵔ ⴰⵍⵡⵙⴰⵏⵉ
@tanhapak1

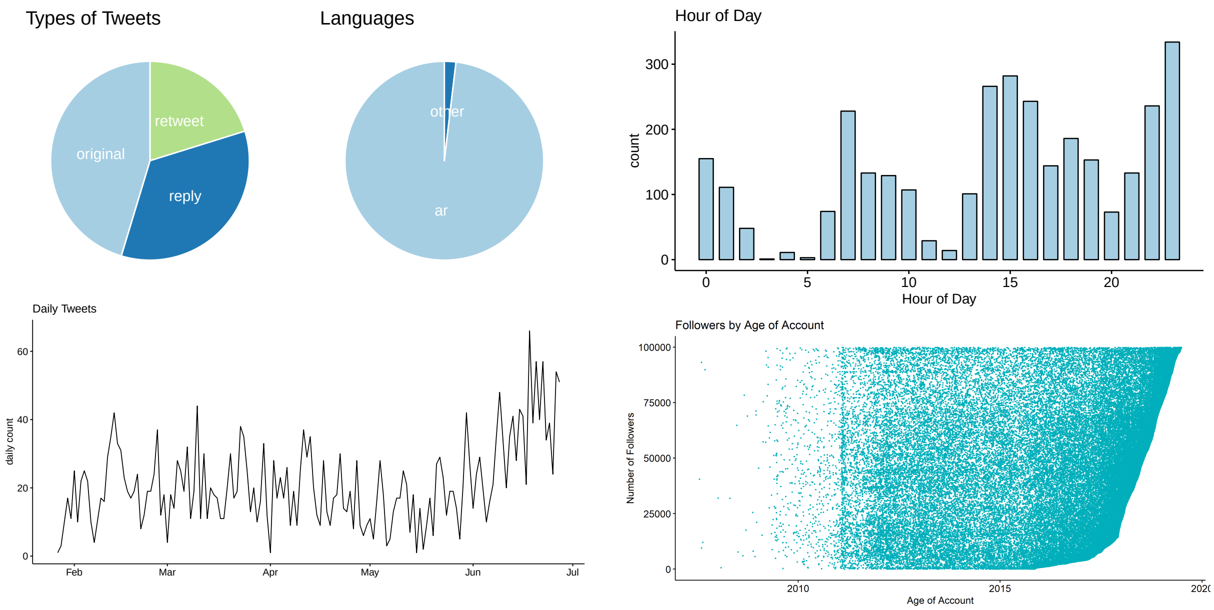
Tweets Tweets & replies Media

Pinned Tweet

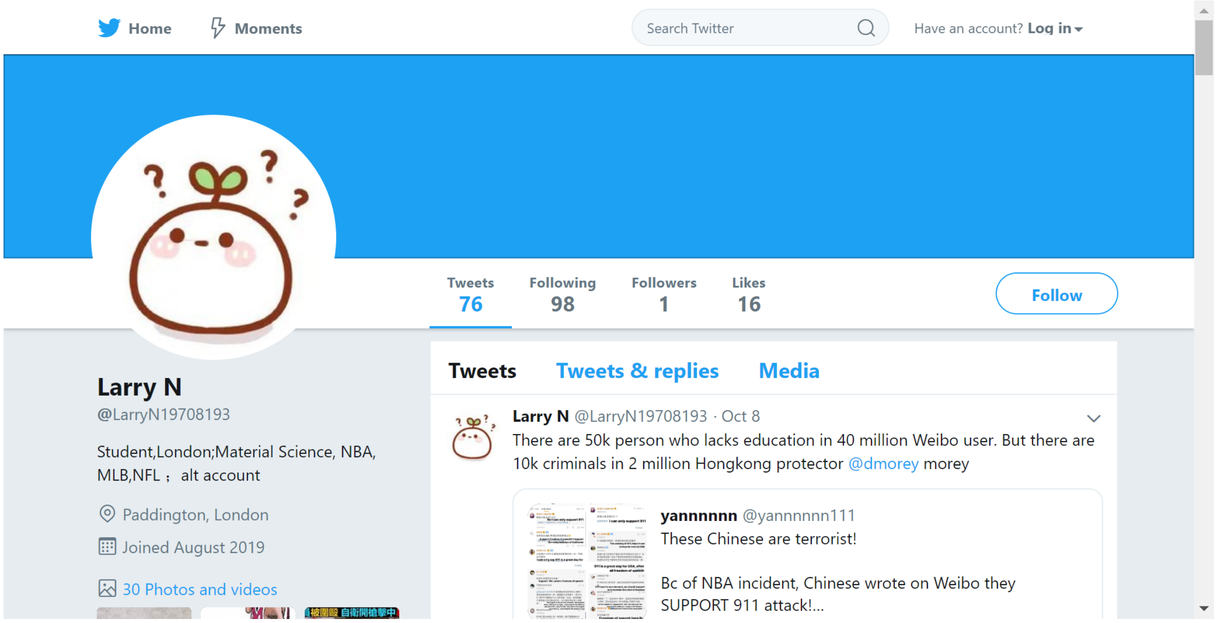
عبير الخولاني🇾🇪ⴰⵏⵉⵔ ⴰⵍⵡⵙⴰⵏⵉ @tanhapak1 · 19 Jan 2017

اقسم بعزة الله انه لايوجد عندنا اغلامن #السعودية الا #اليمن ولايوجد اغلامن #اليمن الا

. **@tanhapak1.** Very popular and influential bot/cyborg in the Yemen conversation.



Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Domain**	**C**
Frd/Fol Ratio	0.032	bot-hunter Tier 0	False	*To whom*	191	mohdsalj	49	youtu.be	
Avg Sentiment (0-1)	0.65	bot-hunter Tier 1	0.6767	*Aden*	112	SaudiDRPY	44	tweepsmap.com	
Dormant Followers	37.4%	bot-hunter Tier 2	0.7088	*Bad*	56	dr_zayedalamri	37	t.co	
Bot Followers	42.1%	bot-hunter Tier 3	0.364	*United Arab Emirates*	53	aloqeliy	35	wtn.sa	
Avg Mentions	1.177	botometer	0.1481	*Hajur*	45	hsom67	35	presidenthadi-gov-ye.info	
Retweets Russia Propaganda	True	Debot	False	*Houthi*	41	malarab1	34	okaz.com.sa	



Tweets 76 Following 98 Followers 1 Likes 16

Larry N
@LarryN19708193

Student,London;Material Science, NBA, MLB,NFL； alt account

Paddington, London
Joined August 2019

30 Photos and videos

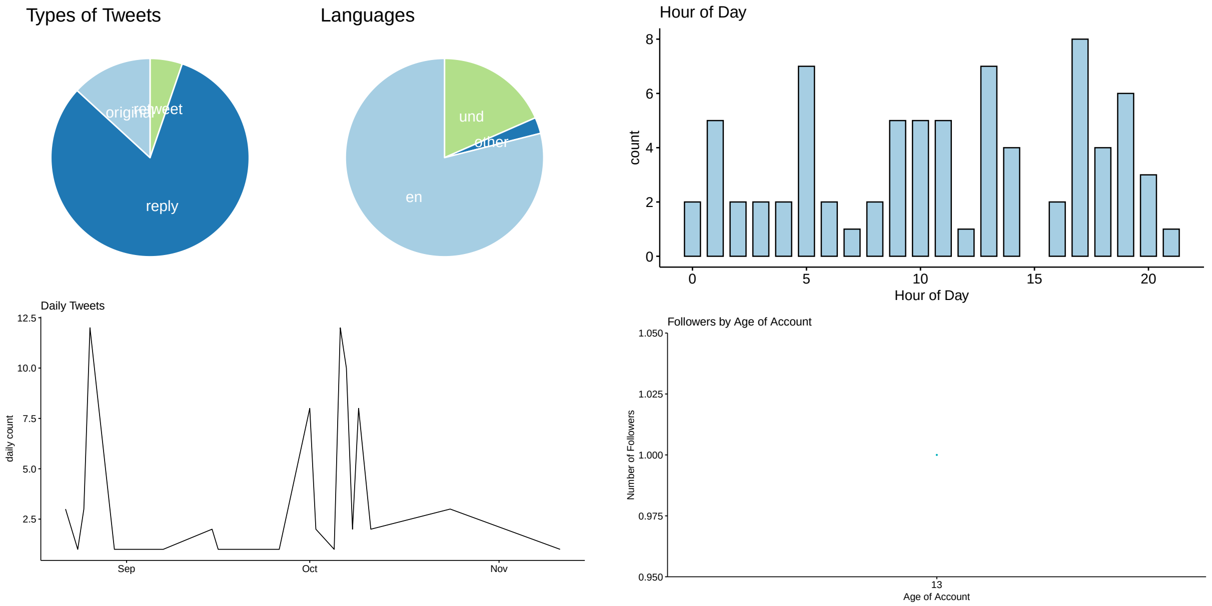
Tweets Tweets & replies Media

Larry N @LarryN19708193 · Oct 8
There are 50k person who lacks education in 40 million Weibo user. But there are 10k criminals in 2 million Hongkong protector @dmorey morey

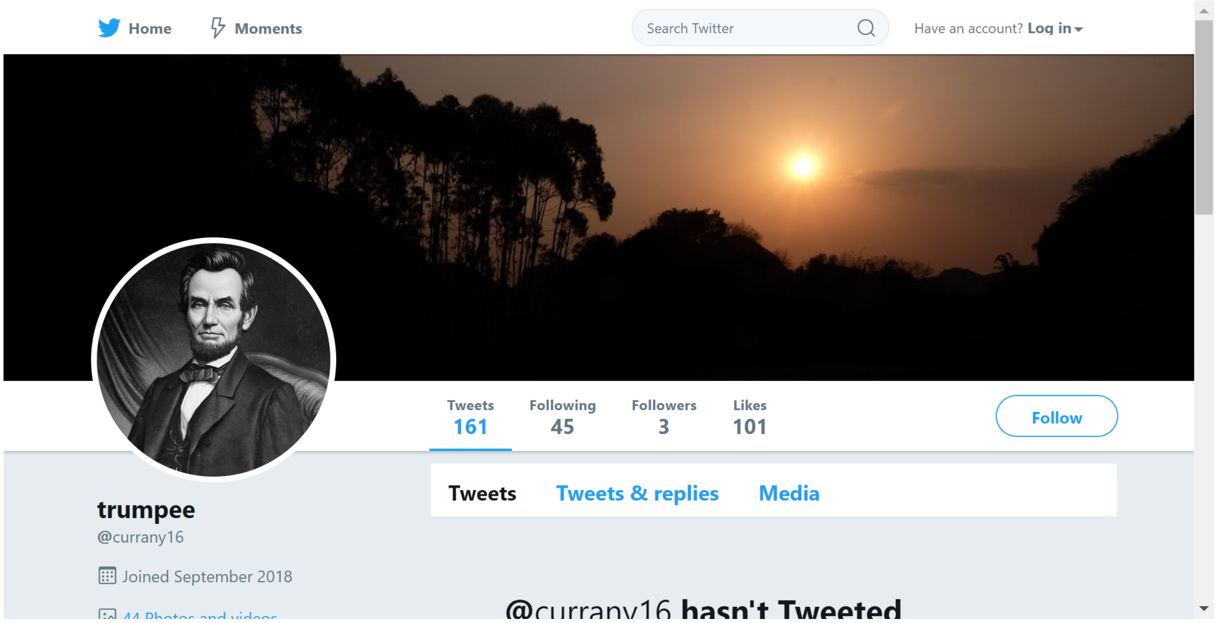
yannnnnn @yannnnnn111
These Chinese are terrorist!

Bc of NBA incident, Chinese wrote on Weibo they SUPPORT 911 attack!...

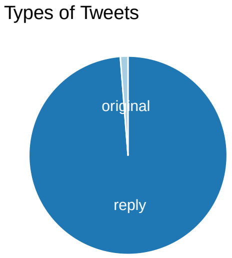
. **@LarryN19708193.** We found @LarryN19708193 linked to the Chinese disinformation campaign against @dmorey and the NBA.

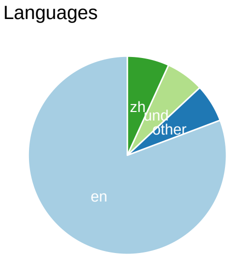


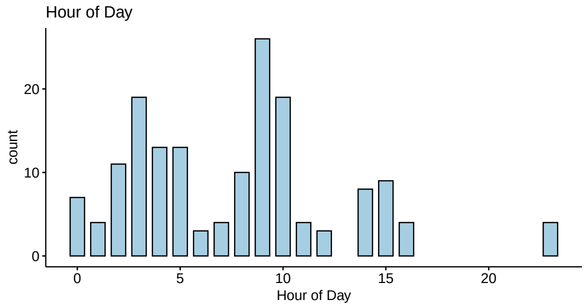
Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	98	bot-hunter Tier 0	False	*Hong Kong Protests*	3	dmorey	29		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.2372	*FreeHongKong*	1	HoustonRockets	8		
Dormant Followers	0%	bot-hunter Tier 2	0.257	*NBABreakdown*	1	CentristReports	7		
Bot Followers	0.0%	bot-hunter Tier 3	0.429	*HongKongProtest*	1	senatemajldr	5		
Avg Mentions	2.188	botometer	0.3184	*TeamDay*	1	alexchang198	5		
Retweets RU Prop.	False	Debot	True	*NBABlockWeek*	1	CriticalCezanne	5		

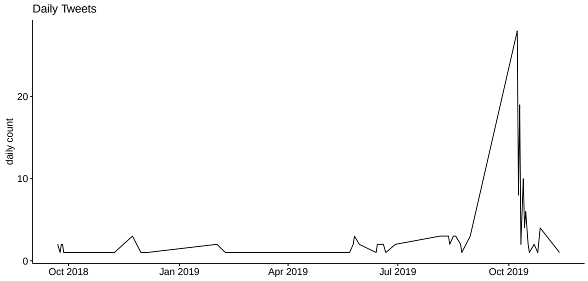


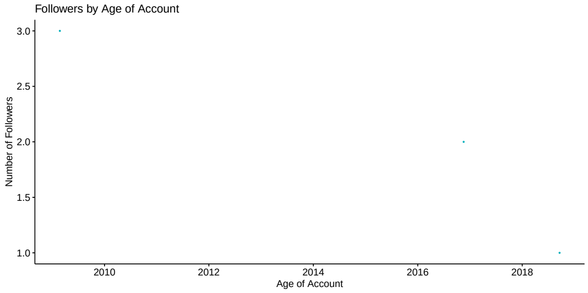
. **@currany16.** We found @currany16 linked to the Chinese effort to attack @dmorey, the NBA, and the United States. In particular, this posted inflammatory memes about Native Americans, Snowden, and other hot button topics.



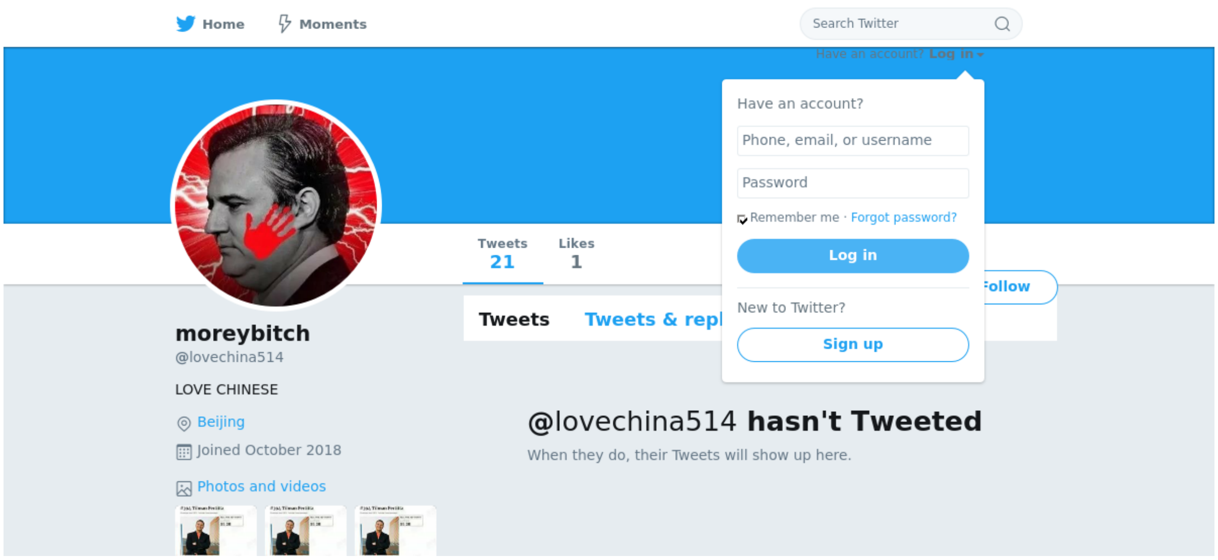








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	15	bot-hunter Tier 0	False	*Opinion*	1.0	NBA	33	Bullshit.You	1.0
Avg Sentiment (0-1)		bot-hunter Tier 1	0.2913	*huaweiforever*	1.0	PDChina	17	dont.it	1.0
Dormant Followers	66.7%	bot-hunter Tier 2	0.4174			dmorey	17	wrong.It	1.0
Bot Followers	0.0%	bot-hunter Tier 3	0.417			realDonaldTrump	13		
Avg Mentions	1.925	botometer	0.2838			TheEconomist	11		
Retweets RU Prop.	False	Debot	True			ObjectFightClub	9		



Home Moments

Search Twitter

Have an account? Log in ▾

Have an account?

Phone, email, or username

Password

☑ Remember me · Forgot password?

Log in

New to Twitter?

Sign up

Tweets 21 Likes 1

Follow

moreybitch
@lovechina514

LOVE CHINESE

Beijing

Joined October 2018

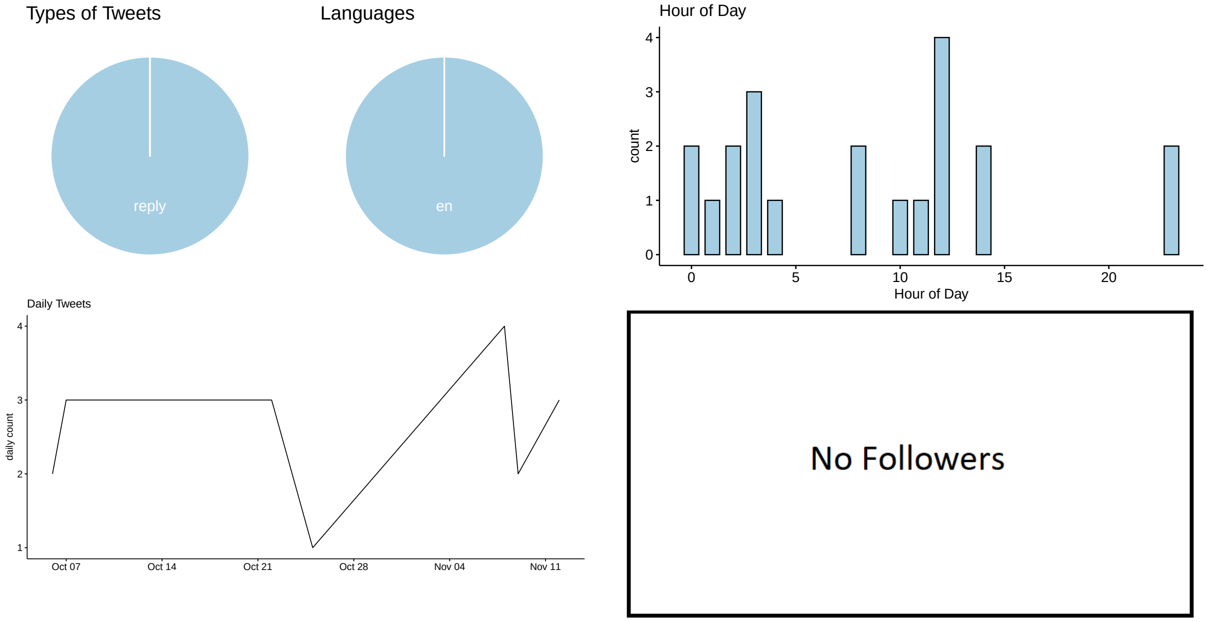
Photos and videos

Tweets Tweets & repl

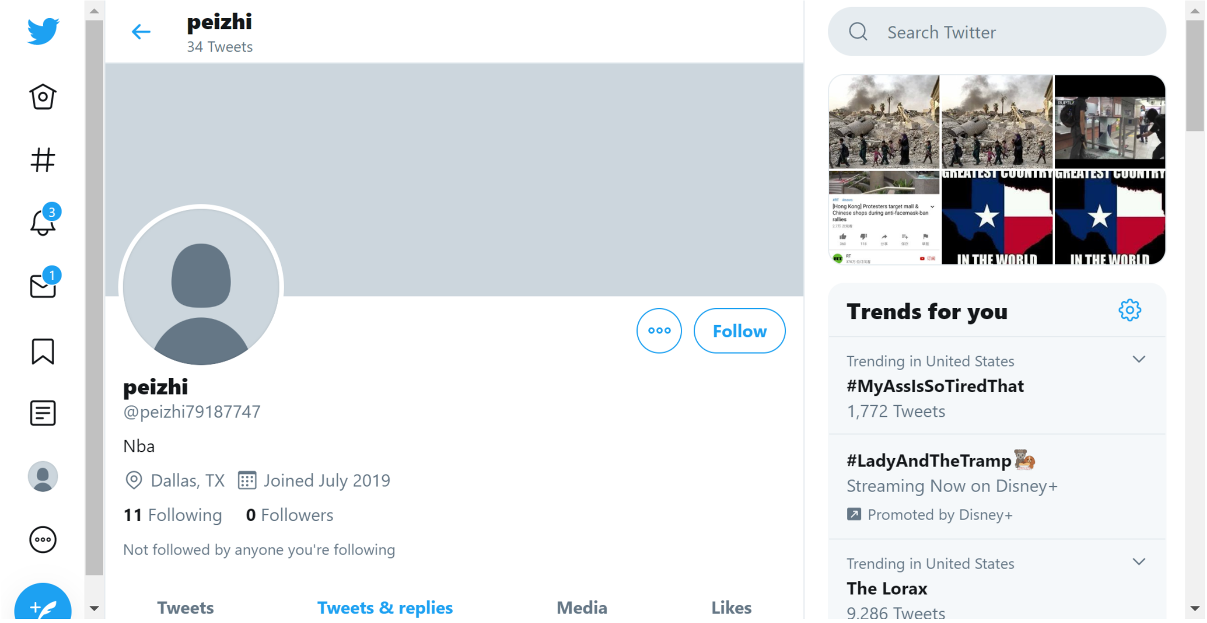
@lovechina514 **hasn't Tweeted**

When they do, their Tweets will show up here.

. **@lovechina514.** We found @lovechina514 linked to Chinese disinformation surrounding @dmorey and the NBA.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio		bot-hunter Tier 0	False			dmorey	13		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.3912			TilmanJFertitta	6		
Dormant Followers		bot-hunter Tier 2	0.3783			HoustonRockets	5		
Bot Followers		bot-hunter Tier 3	0.270			NBA	3		
Avg Mentions	2.238	botometer	0.1821			ToyotaCenter	3		
Retweets RU Prop.	False	Debot	True			espn	2		



peizhi
34 Tweets

Search Twitter

Follow

peizhi
@peizhi79187747

Nba

Dallas, TX Joined July 2019

11 Following 0 Followers

Not followed by anyone you're following

Tweets Tweets & replies Media Likes

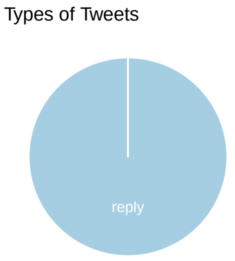
Trends for you

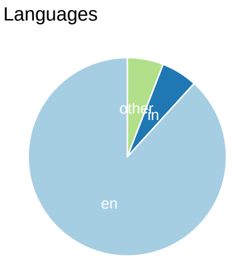
Trending in United States
#MyAssIsSoTiredThat
1,772 Tweets

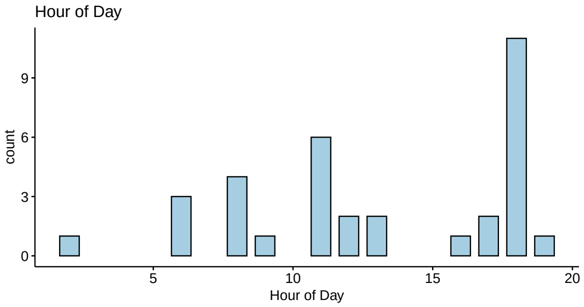
#LadyAndTheTramp
Streaming Now on Disney+
Promoted by Disney+

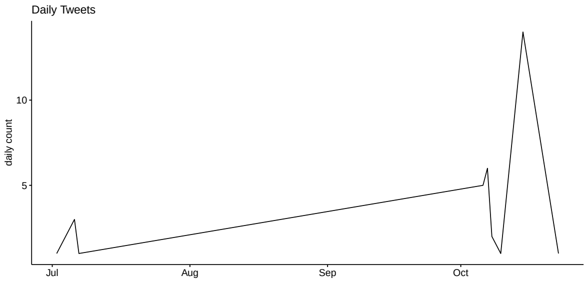
Trending in United States
The Lorax
9,286 Tweets

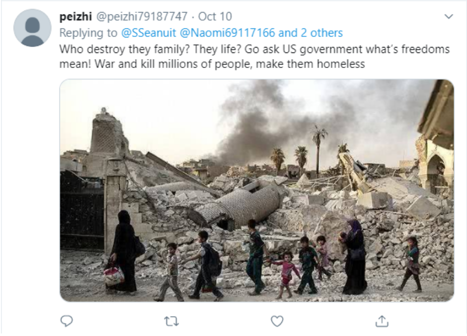
. **@peizhi79187747.** We found @peizhi79187747 linked to China disinformation efforts against @dmorey and the NBA.











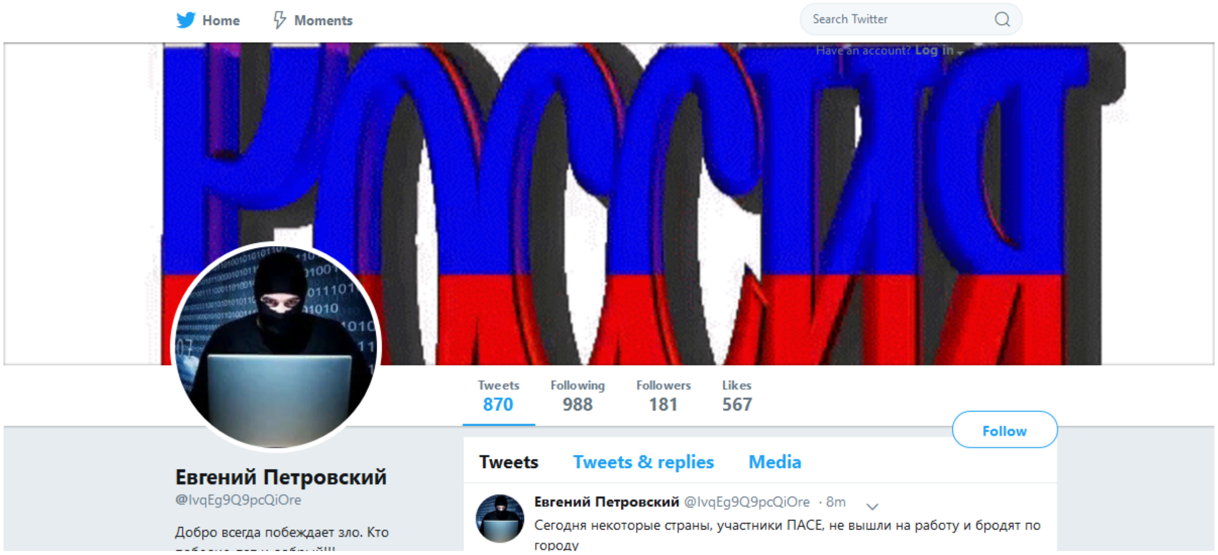
peizhi @peizhi79187747 · Oct 10
Replying to @SSeanuit @Naomi69117166 and 2 others
Who destroy they family? They life? Go ask US government what's freedoms mean! War and kill millions of people, make them homeless

Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio		bot-hunter Tier 0	False			KingJames	13		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.277			dmorey	11		
Dormant Followers		bot-hunter Tier 2	0.3395			NBA	7		
Bot Followers		bot-hunter Tier 3	0.265			SolomonYue	6		
Avg Mentions	2.618	botometer	0.3552			Yanting_Huang	3		
Retweets RU Prop.	False	Debot	True			wojespn	3		

E.11 Random String Screen Name Accounts

We've found many accounts that contain a 15 digit randomly generated alpha-numeric string. We've developed machine learning methods to detect these strings, and have applied it to large streams of data, identifying millions of bot accounts that have this feature. There are many types of bot accounts that have this feature.

In the next few pages we've given some examples of different types of bot accounts that all have a randomly generated alpha-numeric string for a screen_name.

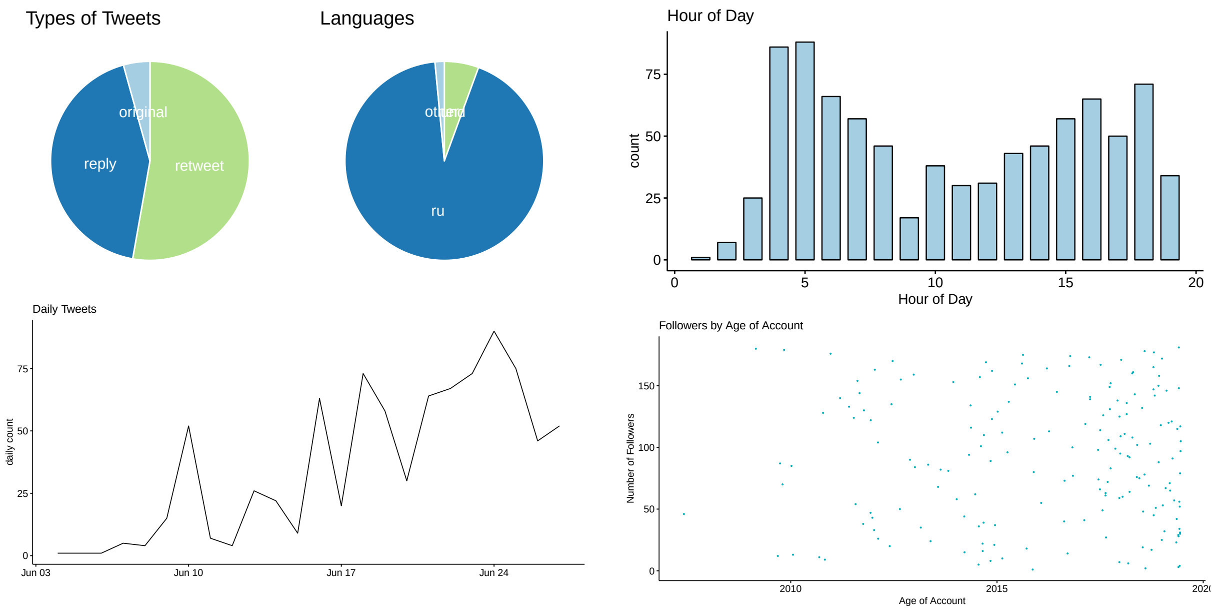


Евгений Петровский
@IvqEg9Q9pcQiOre

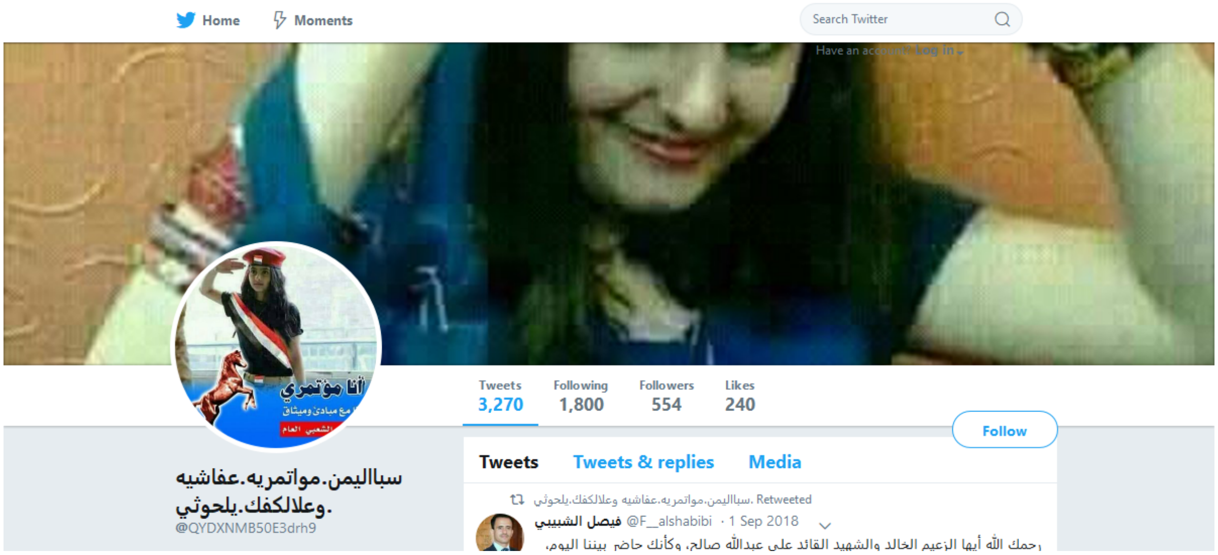
Добро всегда побеждает зло. Кто

Евгений Петровский @IvqEg9Q9pcQiOre · 8m
Сегодня некоторые страны, участники ПАСЕ, не вышли на работу и бродят по городу

. **@IvqEg9Q9pcQiOre.** We found @IvqEg9Q9pcQiOre to be a pro-Russia and anti-Ukraine bot.



Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Domain**	**Count**
Frd/Fol Ratio	5.462	bot-hunter Tier 0	True	*Sports*	1	Dialog$_{U}A$	42	youtube.com	9
Avg Sentiment (0-1)	0.662	bot-hunter Tier 1	0.5521	*in404*	1	R8yEnDsAA1D92Dd	32	inforuss.info	3
Dormant Followers	1.1%	bot-hunter Tier 2	0.4151	*On the mark*	1	bob20006	31	faleev.net	3
Bot Followers	26.9%	bot-hunter Tier 3	0.334	*Db*	1	prof$_{p}reobr$	26	ren.tv	2
Avg Mentions	1.265	botometer	0.2515	*Boyarsky*	1	dolg132005	24	ria.ru	2
Retweets Russia Propaganda	False	Debot	False	*truce*	1	RussiansForward	23	ukraina.ru	2

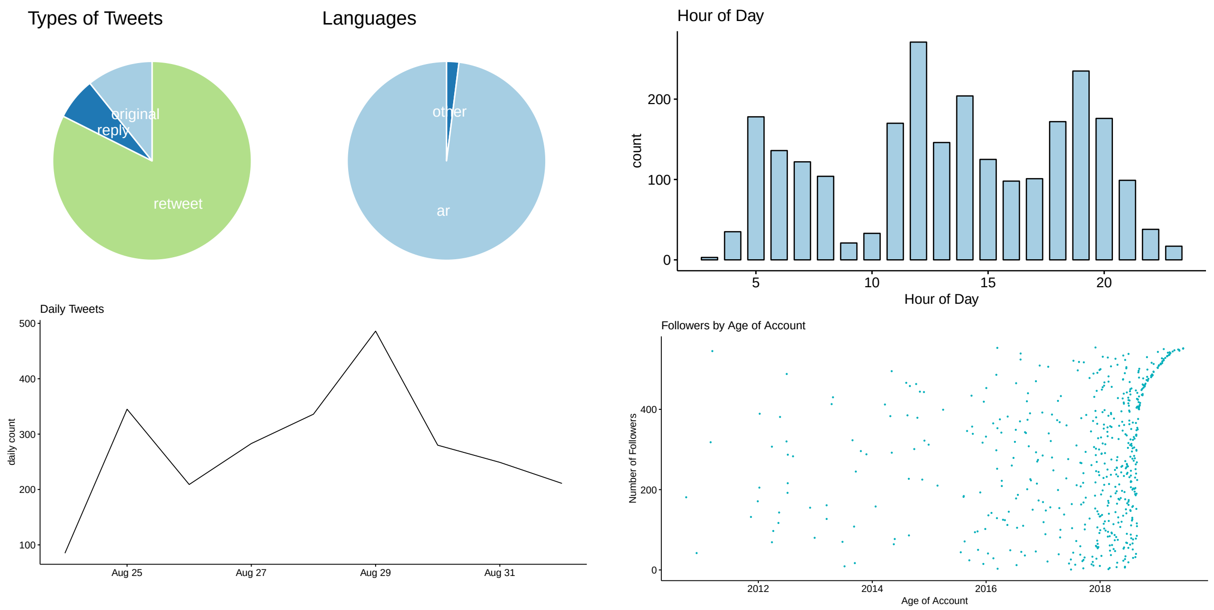


Tweets 3,270 Following 1,800 Followers 554 Likes 240

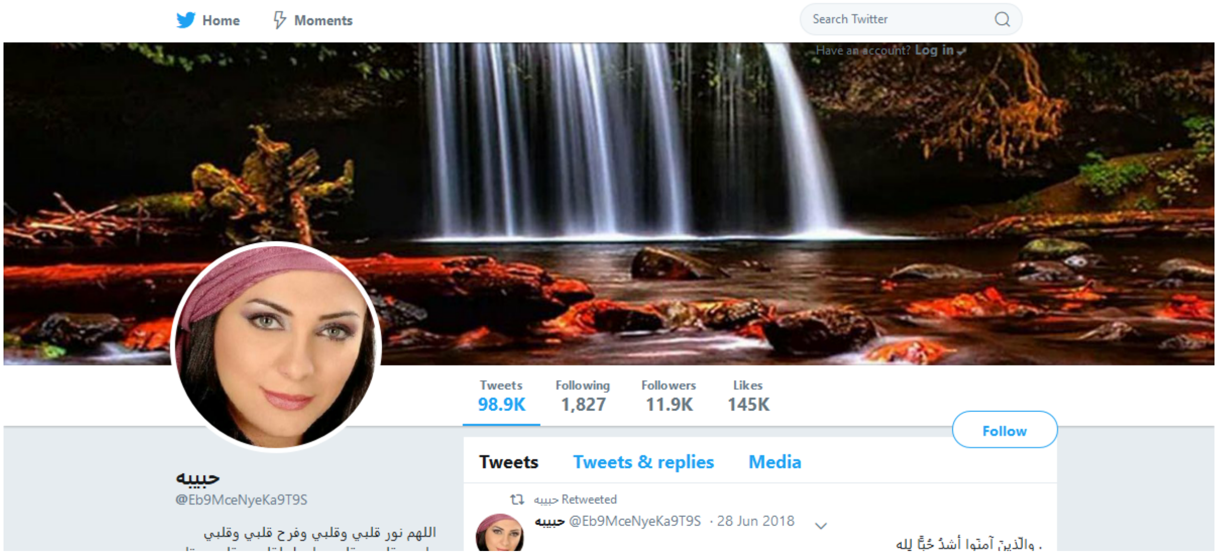
سباليمن.مواتمريه.عفاشيه
وعلالكفك.يلحوثي.
@QYDXNMB50E3drh9

سباليمن.مواتمريه.عفاشيه وعلالكفك.يلحوثي. Retweeted
فيصل الشبيبي @F__alshabibi · 1 Sep 2018
رحمك الله أيها الزعيم الخالد والشهيد القائد على عبدالله صالح، وكأنك حاضر بيننا اليوم،

. **@QYDXNMB50E3drh9.** We found @QYDXNMB50E3drh9 linked to the conflict in Yemen



Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Domain**	**Count**
Frd/Fol Ratio	3.251	bot-hunter Tier 0	True	*To whom*	47	RepublicanYemen	54	jan14news.net	16
Avg Sentiment (0-1)	0.71	bot-hunter Tier 1	0.331	*Houthi*	42	tareek20015	44	youtu.be	13
Dormant Followers	28.9%	bot-hunter Tier 2	0.4499	*Sana'a*	33	asas12367	29	khabaragency.net	8
Bot Followers	22.6%	bot-hunter Tier 3	0.314	*Hodeidah*	25	Ben$_{A}fash$	27	2dec.net	6
Avg Mentions	1.106	botometer	0.4337	*Saada*	18	SAlghobari	27	marebtoday.net	4
Retweets Russia Propaganda	True	Debot	False	*Taiz*	17	feel8queen	26	pscp.tv	4



حبيبه
@Eb9MceNyeKa9T9S

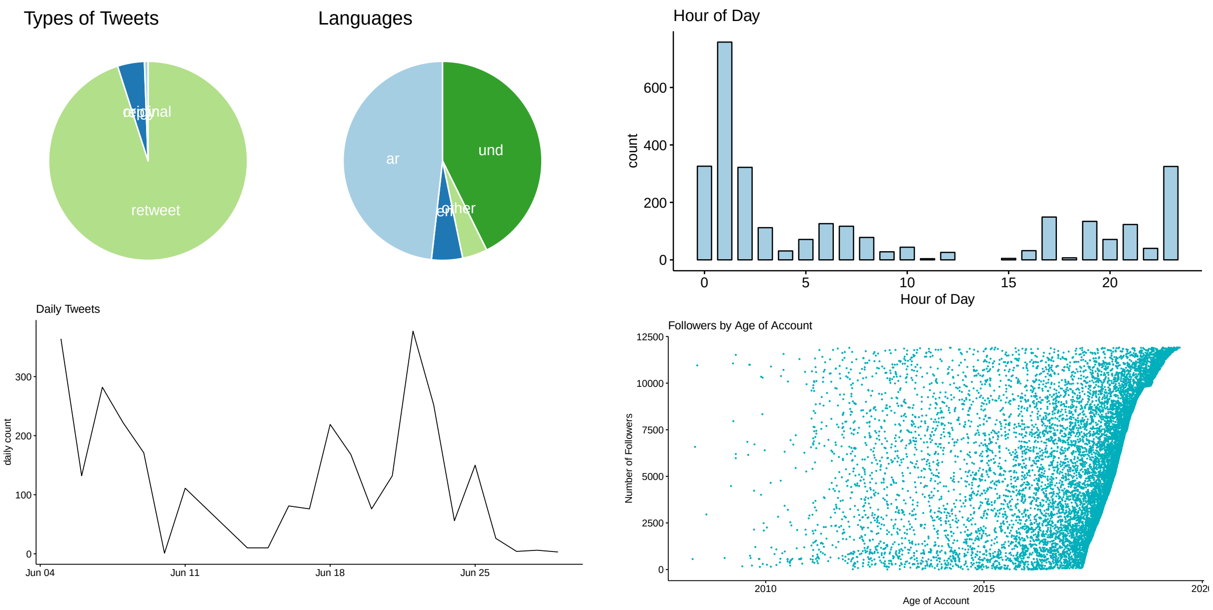
اللهم نور قلبي وقلبي وفرح قلبي وقلبي

حبيبه Retweeted

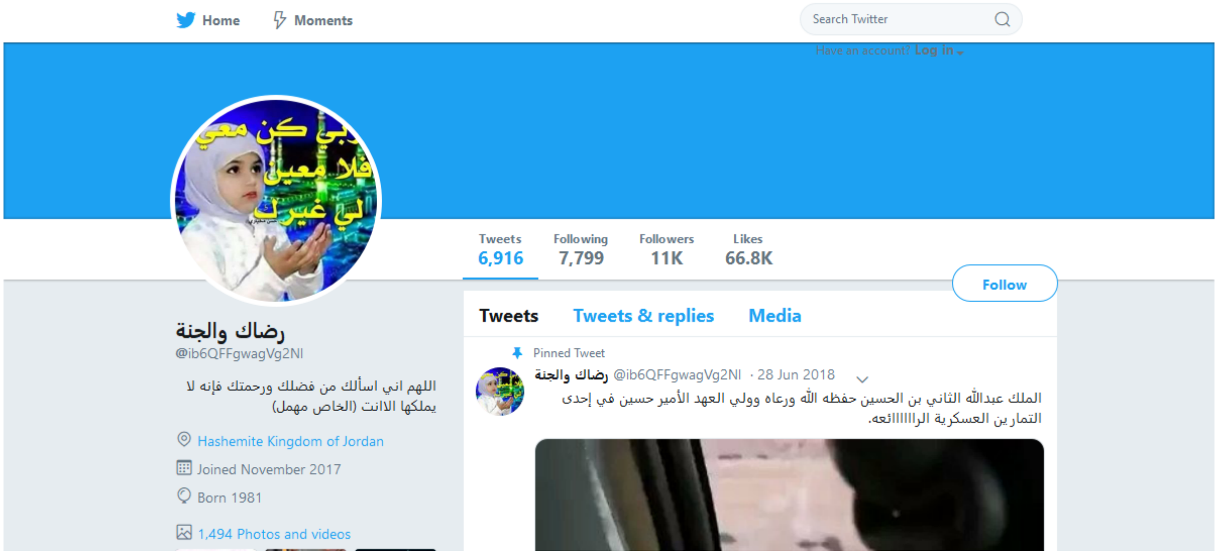
حبيبه @Eb9MceNyeKa9T9S · 28 Jun 2018

والذين آمنوا أشدُّ حبًّا لله .

. @Eb9MceNyeKa9T9S. This is a popular Arabic bot.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Cc
Frd/Fol Ratio	0.153	bot-hunter Tier 0	True	*Cry mazlum*	6	Eb9MceNyeKa9T9S	729	du3a.org	
Avg Sentiment (0-1)	0.694	bot-hunter Tier 1	0.6032	*Saleh Al-Ghamdi*	6	Blanca_Garza3	356	d3waapp.org	
Dormant Followers	41.5%	bot-hunter Tier 2	0.5492	*Response time*	6	Alfredochiribo1	323	8uran.online	
Bot Followers	25.7%	bot-hunter Tier 3	0.469	*Riyadh*	5	arunsuikar	263	AthanTweets.com	
Avg Mentions	2.663	botometer	0.5768	*Al-Aqsa*	5	MohamedAl_Adawi	244	zad-muslim.com	
Retweets Russia Propaganda	False	Debot	False	*Win Yavonen with good*	4	AbdehHalim	225	apple.com	


Home Moments

Search Twitter

Have an account? Log in

Tweets 6,916 | Following 7,799 | Followers 11K | Likes 66.8K

Follow

Tweets | Tweets & replies | Media

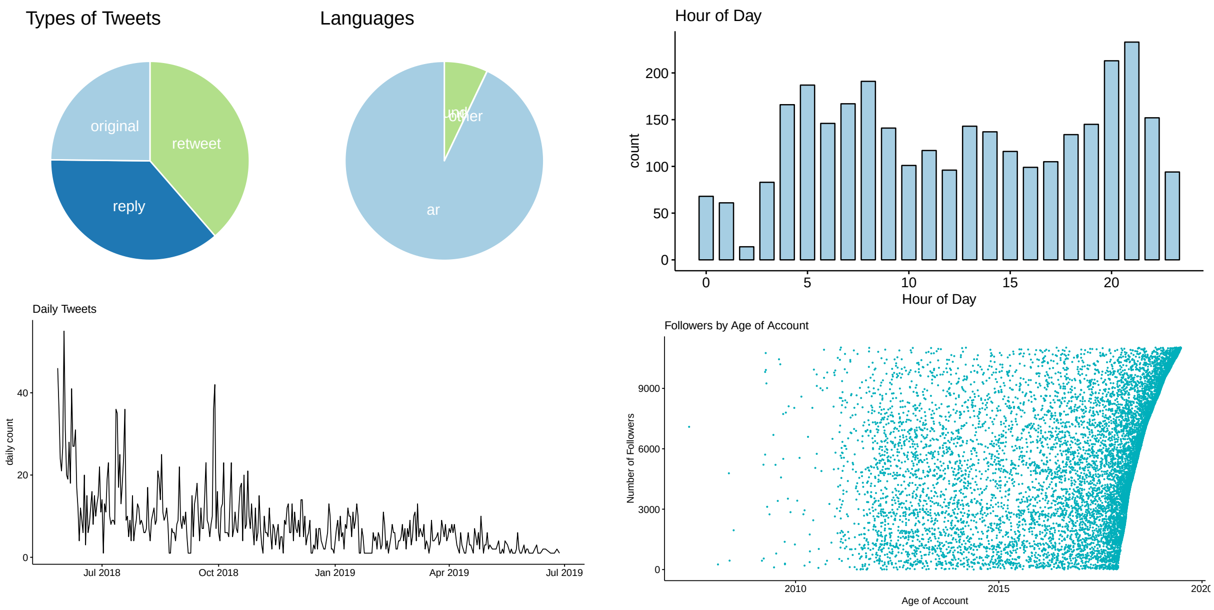
رضاك والجنة
@ib6QFFgwagVg2Nl

اللهم اني اسألك من فضلك ورحمتك فإنه لا
يملكها الانت (الخاص مهمل)

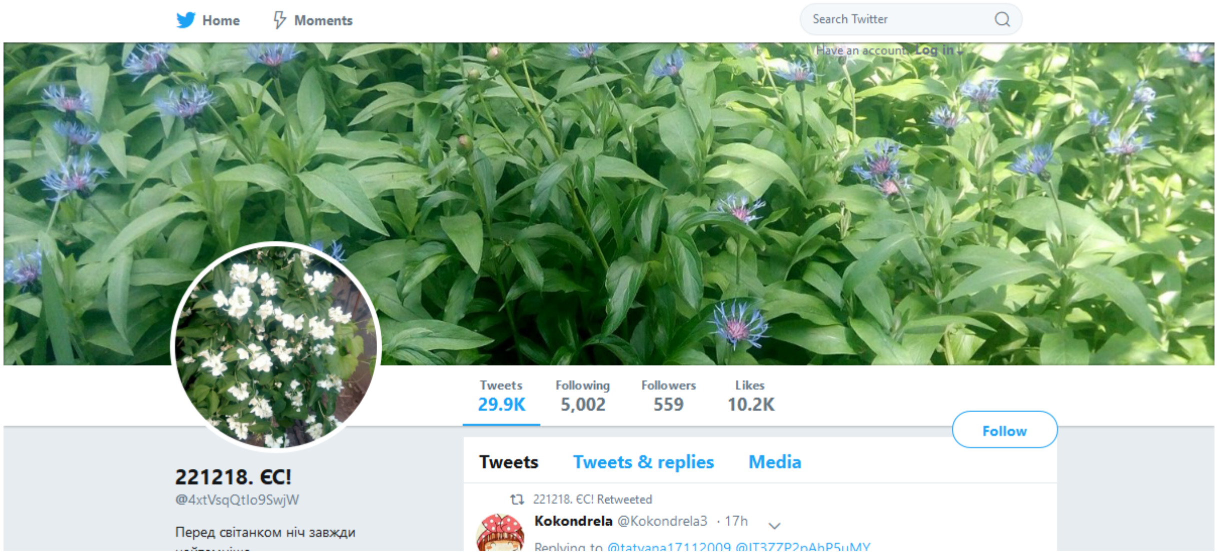
Hashemite Kingdom of Jordan
Joined November 2017
Born 1981
1,494 Photos and videos

Pinned Tweet
رضاك والجنة @ib6QFFgwagVg2Nl · 28 Jun 2018
الملك عبدالله الثاني بن الحسين حفظه الله ورعاه وولي العهد الأمير حسين في إحدى
التمارين العسكرية الراااااائعه.

. **@ib6QFFgwagVg2Nl.** We found @ib6QFFgwagVg2Nl to be a popular Arabic propaganda bot



Metric	Value	**Model**	**Probability**	**Hashtag**	**Count**	**Mention**	**Count**	**Domain**
Frd/Fol Ratio	0.707	bot-hunter Tier 0	True	*personal opinion*	11	ib6QFFgwagVg2Nl	219	du3a.org
Avg Sentiment (0-1)	0.655	bot-hunter Tier 1	0.4661	*Jordan*	8	dzdalm	133	jo24.net
Dormant Followers	26.5%	bot-hunter Tier 2	0.6267	*Eid_Halawiyat_Rekn_Alqasim 5*	6	AaAa0531420	120	royanews.tv
Bot Followers	17.9%	bot-hunter Tier 3	0.581	*Response time*	6	dot_amar	72	athkarapp.online
Avg Mentions	1.194	botometer	0.4745	*good morning*	5	abdalhkeem571	63	ksu.edu.sa
Retweets Russia Propaganda	False	Debot	False	*Jordan*	5	badr_s_alrajhi	45	soundcloud.com



Home   Moments   Search Twitter

Have an account? Log in

Tweets 29.9K   Following 5,002   Followers 559   Likes 10.2K

Follow

221218. ЄС!

@4xtVsqQtIo9SwjW

Перед світанком ніч завжди

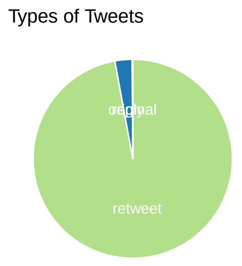
Tweets   Tweets & replies   Media

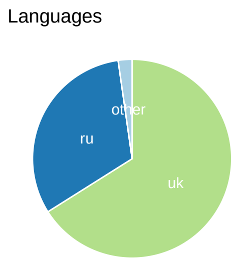
221218. ЄС! Retweeted

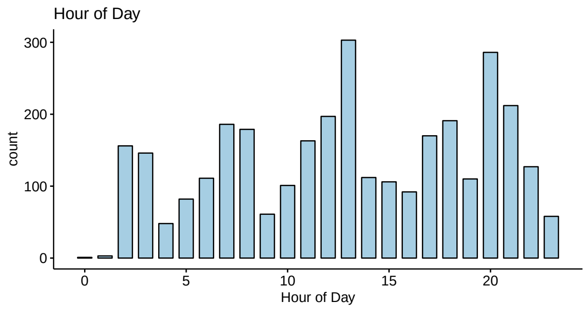
Kokondrela @Kokondrela3 · 17h

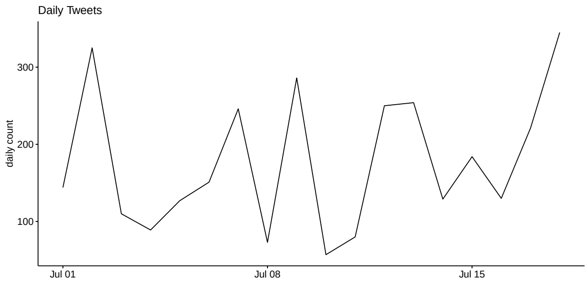
Replying to @tatyana17112009 @IT3Z7P2nAhP5uMY

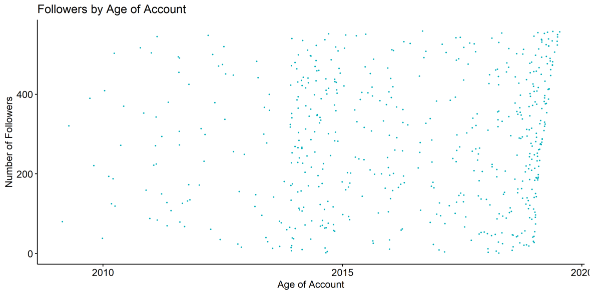
. @4xtVsqQtIo9SwjW. We found @4xtVsqQtIo9SwjW linked to information operations in Ukraine. The ORA Twitter Report has this account as one of the top "Other Influencers" in conversations around the Ukrainian President.



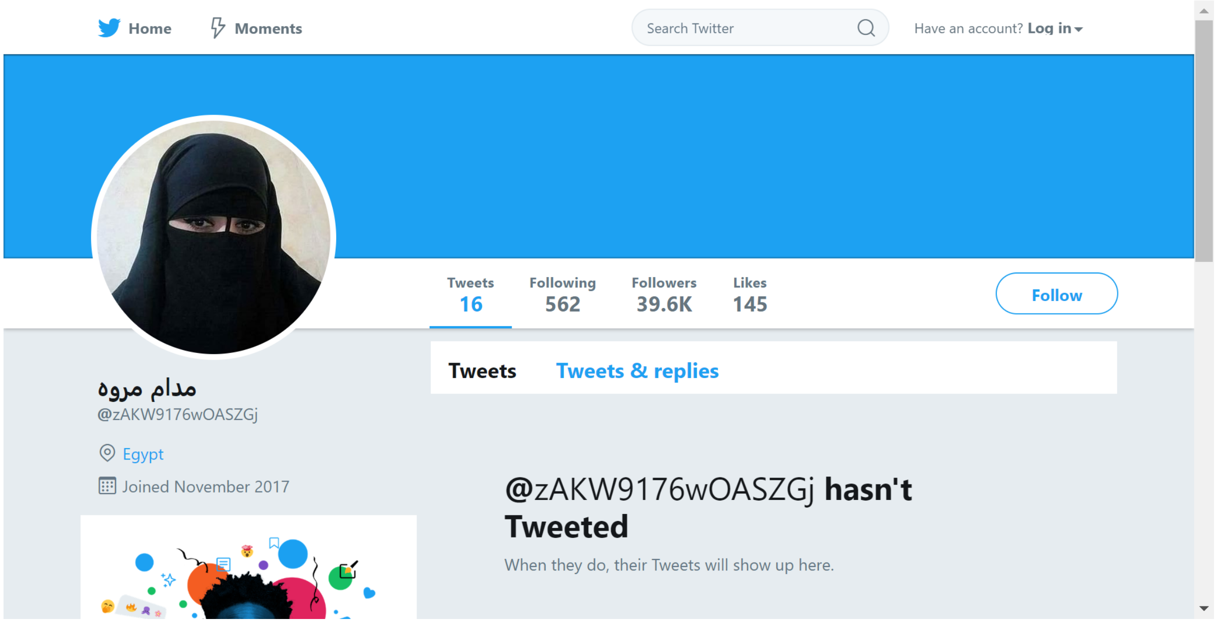








Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	8.946	bot-hunter Tier 0	False	*news*	5	APUkraine	173	youtu.be	12
Avg Sentiment (0-1)		bot-hunter Tier 1	0.483	*news*	4	FashDonetsk	112	newsua.one	7
Dormant Followers	2.7%	bot-hunter Tier 2	0.6127	*SBU*	2	4xtVsqQtIo9SwjW	83	lviv-company.in.ua	6
Bot Followers	20.6%	bot-hunter Tier 3	0.361	*Our_parad2019*	2	tatyana17112009	77	nizhyn.in.ua	5
Avg Mentions	2.164	botometer	0.4136	*FreeSentsov*	2	UkrBereza	73	espreso.tv	5
Retweets RU Prop.	True	Debot	False	*Language is uniting*	2	PrytSLU	69	facebook.com	5



Tweets 16 Following 562 Followers 39.6K Likes 145

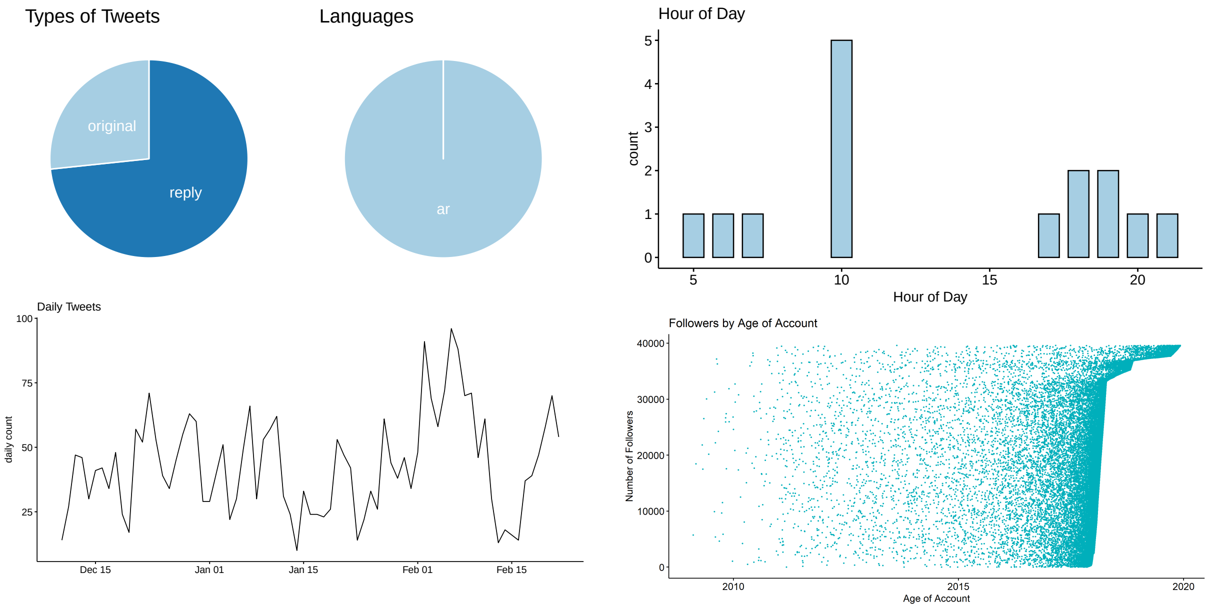
مدام مروه
@zAKW9176wOASZGj

Egypt
Joined November 2017

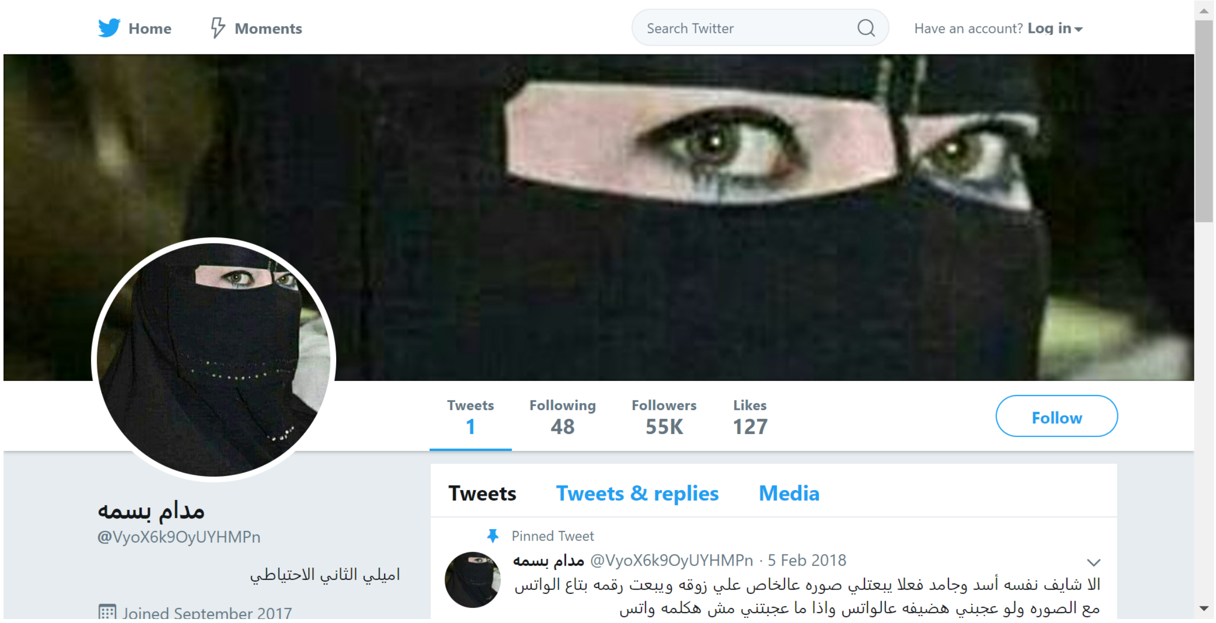
@zAKW9176wOASZGj hasn't Tweeted

When they do, their Tweets will show up here.

. **@zAKW9176wOASZGj.** We found @zAKW9176wOASZGj linked to intimidation accounts.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.014	bot-hunter Tier 0	False			9lkAxYE1sn4Kvyv	5		
Avg Sentiment (0-1)		bot-hunter Tier 1	0.64			soso56o67	2		
Dormant Followers	72.4%	bot-hunter Tier 2	0.6686			gEHS8N0m1ZHrfwh	1		
Bot Followers	55.4%	bot-hunter Tier 3	0.374			1i5uamlq09RpCbF	1		
Avg Mentions	1	botometer	0.5768			salwa11442182	1		
Retweets RU Prop.	False	Debot	True			VyoX6k9OyUYHMPn	1		



Tweets 1 · Following 48 · Followers 55K · Likes 127

Tweets · Tweets & replies · Media

مدام بسمه
@VyoX6k9OyUYHMPn

اميلي الثاني الاحتياطي

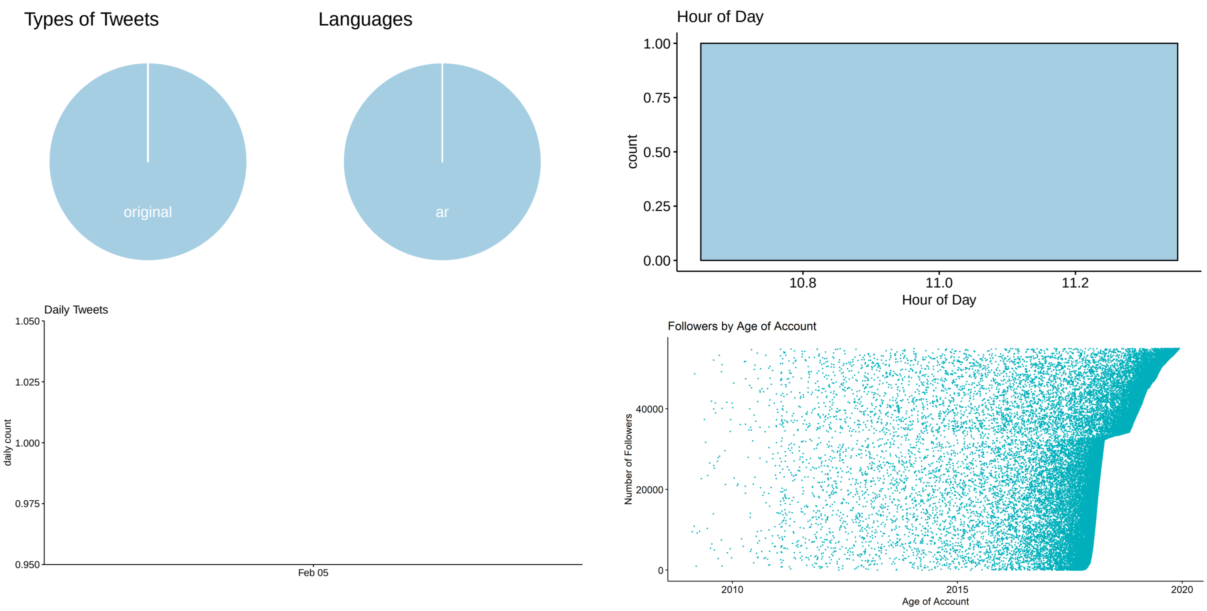
Joined September 2017

Pinned Tweet

مدام بسمه @VyoX6k9OyUYHMPn · 5 Feb 2018

الا شايف نفسه أسد وجامد فعلا يبعتلي صوره عالخاص علي زوقه ويبعت رقمه بتاع الواتس
مع الصوره ولو عجبني هضيفه عالواتس واذا ما عجبتني مش هكلمه واتس

. **@VyoX6k9OyUYHMPn.** We found @VyoX6k9OyUYHMPn linked to intimidation attacks.



Metric	Value	Model	Probability	Hashtag	Count	Mention	Count	Domain	Count
Frd/Fol Ratio	0.001	bot-hunter Tier 0	False						
Avg Sentiment (0-1)		bot-hunter Tier 1	0.4883						
Dormant Followers	62.4%	bot-hunter Tier 2	NA						
Bot Followers	52.0%	bot-hunter Tier 3	0.470						
Avg Mentions		botometer	0.9726						
Retweets RU Prop.	False	Debot	True						

Bibliography

[1] Alexander Abad-Santos. How memes became the best weapon against chinese internet censorship - the atlantic. https://www.theatlantic.com/international/archive/2013/06/how-memes-became-best-weapon-against-chinese-internet-censorship/314618/, June 2013. (Accessed on 04/06/2019). 5.5

[2] Lawrence Abrams. Beware of fake john mcafee and tesla cryptocurrency giveaways. https://www.bleepingcomputer.com/news/security/beware-of-fake-john-mcafee-and-tesla-cryptocurrency-giveaways/, June 2019. (Accessed on 12/11/2019). 6.4.2

[3] Steve Abrams. Beyond propaganda: Soviet active measures in putin's russia. *Connections*, 15(1):5–31, 2016. 1.3

[4] Noor Abu-El-Rub and Abdullah Mueen. Botcamp: Bot-driven interactions in social campaigns. In *The World Wide Web Conference*, pages 2529–2535. ACM, 2019. 6.2

[5] Russell L Ackoff. From data to wisdom. *Journal of applied systems analysis*, 16(1):3–9, 1989. 1.2

[6] Kayode Sakariyah Adewole, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak. Malicious accounts: dark of the social networks. *Journal of Network and Computer Applications*, 79:41–67, 2017. 3.2.1, 3.2.1, 3.2.4

[7] Central Intelligence Agency. Romeo spies — central intelligence agency. https://www.cia.gov/news-information/featured-story-archive/2018-featured-story-archive/romeo-spies.html, February 2018. (Accessed on 12/16/2019). 6.5.1

[8] Swedish Defence Reserch Agency. Antalet botar på twitter ökar inför valet - totalförsvarets forskningsinstitut. https://www.foi.se/press--nyheter/nyheter/nyhetsarkiv/2018-08-29-antalet-botar-pa-twitter-okar-infor-valet.html, 08 2018. (Accessed on 11/20/2018). 3.8

[9] Nesreen K Ahmed, Ryan Rossi, John Boaz Lee, Theodore L Willke, Rong Zhou, Xiangnan Kong, and Hoda Eldardiry. Learning role-based graph embeddings. *arXiv preprint arXiv:1802.02896*, 2018. 4.2.2

[10] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. People are strange when you're a stranger: Impact and influence of bots on social networks. *Links*, 697(483,151):1–566, 2012. 3.2.4

[11] Media Ajir and Bethany Vailliant. Russian information warfare: Implications for deterrence theory. *Strategic Studies Quarterly*, 12(3):70–89, 2018. 1.3, 1.7

[12] David S Alberts. Defensive information warfare. Technical report, NATIONAL DEFENSE UNIV WASHINGTON DC INST FOR NATIONAL STRATEGIC STUDIES, 1996. 7.2

[13] Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. Detect me if you can: Spam bot detection using inductive representation learning. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 148–153. ACM, 2019. 4.2.4

[14] Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 15(5): 475–491, 2016. 3.2, 3.2.4

[15] Priscilla Alvarez and Taylor Hosking. The full text of mueller's indictment of 13 russians'. *The Atlantic, 16th February*, 2018. A.2.1

[16] A. Ananthalakshmi. Ahead of malaysian polls, bots flood twitter with pro-government..., Apr 2018. URL `https://www.reuters.com/article/us-malaysia-election-socialmedia/ahead-of-malaysian-polls-bots-flood-twitter-with-pro-government-messages-idUSKBN1HR2AQ`. 3.1

[17] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*, 2018. 4.2.1

[18] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*, 2018. 4.2.1

[19] Matthew Babcock, David M Beskow, and Kathleen M Carley. Beaten up on twitter? exploring fake news and satirical responses during the black panther movie event. In Robert Thompson, Christopher Dancy, Ayaz Hyder, and Halil Bisgin, editors, *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, volume 10899 of *Lecture Notes in Computer Science*, pages 97–103, Cham, Switzerland, 2018. Springer. B.3.6

[20] Adam Badawy, Aseel Addawood, Kristina Lerman, and Emilio Ferrara. Characterizing the 2016 russian ira influence campaign. *Social Network Analysis and Mining*, 9(1):31, 2019. 6.2, A.2.1

[21] David A Bader, Shiva Kintali, Kamesh Madduri, and Milena Mihail. Approximating betweenness centrality. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 124–137. Springer, 2007. 3.2.6

[22] Charles K Bartles. Getting gerasimov right. *Military Review*, 96(1):30–38, 2016. 1.3, 1.5

[23] Christian Bauckhage, Kristian Kersting, and Fabian Hadiji. Mathematical models of fads explain the temporal dynamics of internet memes. In *ICWSM*, pages 22–30, 2013. 5.2.1,

5.2.3

[24] Robert F. Baumann. A central asian perspective on russian soft power: The view from tashkent. *Military Review*, 98(4):50–63, 2018. 1.6.2

[25] Alex Bavelas. A mathematical model for group structures. *Human organization*, 7(3): 16–30, 1948. 3.2.6

[26] Matthew Benigni and Kathleen M Carley. From tweets to intelligence: Understanding the islamic jihad supporting community on twitter. In *Social, Cultural, and Behavioral Modeling: 9th International Conference, SBP-BRiMS 2016, Washington, DC, USA, June 28-July 1, 2016, Proceedings 9*, pages 346–355. Springer, 2016. 2.2, 2.3, 3.1, 3.2

[27] Matthew Benigni, Kenneth Joseph, and Kathleen M Carley. Mining online communities to inform strategic messaging: practical methods to identify community-level insights. *Computational and Mathematical Organization Theory*, 24(2):224–242, 2018. 6.5.1, 7.3

[28] Matthew C Benigni, Kenneth Joseph, and Kathleen M Carley. Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter. *PloS one*, 12(12):e0181405, 2017. 3.2.4, 3

[29] Jon Louis Bentley. A survey of techniques for fixed radius near neighbor searching. Technical report, 1975. 5.4.2

[30] David Beskow and Kathleen M Carley. Introducing bothunter: A tiered approach to detection and characterizing automated activity on twitter. In Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson, editors, *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018. (document), 3.1, 3.2.1, 3.1, 3.2.7, 3.3.4, 3.3, 3.5, 4.6, 4.7, 5.4.3, 6.2, 6.4.1, 6.5.3, A.4.5, B.1, B.5

[31] David Beskow and Kathleen M Carley. Using random string classification to filter and annotate automated accounts. In Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson, editors, *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018. 3.1, 3.2.1, 3.2, 3.3.4, 3.3, C.2.1

[32] David Beskow and Kathleen M. Carley. You are known by your friends: Leveraging network metrics for bot detection in twitter. In *Open Source Intelligence and Cyber Crime - Social Media Analytics*. Springer, 2020. 2.2, 4.1

[33] David M Beskow and Kathleen M Carley. Bot conversations are different: Leveraging network metrics for bot detection in twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 825–832. IEEE, 2018. (document), 3.1, 3.1.1, 3.3, 3.4, 3.4, 3.4.2, 4.6, A.4.5

[34] David M. Beskow and Kathleen M. Carley. Social cybersecurity: An emerging national security requirement. *Military Review*, 99(2):117–127, 2019. 1.1, 2, A.4.5, B.1, B.2.4, B.4.1, B.4.2

[35] David M Beskow and Kathleen M. Carley. Army must regain initiative in social cyberwar. *Army Magazine*, 69(8):24–28, 2019. 1.1, 2, A.6

[36] David M Beskow and Kathleen M Carley. Characterization and comparison of russian and chinese disinformation campaigns. In Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu, editors, *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, Lecture Notes in Social Networks. Springer International Publishing, New York, 2020. URL `https://www.springer.com/gp/book/9783030426989`. 4.3.3, 6.5.2

[37] David M Beskow and Kathleen M Carley. You are known by your friends: Leveraging network metrics for bot detection in twitter. In Mohammad Taybei, Uwe Glässer, and David Skillicorn, editors, *Open Source Intelligence and Security Informatics*, chapter 3. Springer, New York, 2020. 3.1, 4.6, 6.5.2

[38] David M. Beskow, Sumeet Kumar, and Kathleen M. Carley. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170, mar 2020. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102170. URL `http://www.sciencedirect.com/science/article/pii/S0306457319307988`. 6.5.2

[39] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11-7), 2016. 3.1, 6.2, B.1

[40] Sajid Yousuf Bhat and Muhammad Abulaish. Community-based features for identifying spammers in online social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 100–107. IEEE, 2013. 3.2, 3.2.4

[41] John Biersack and Shannon O'lear. The geopolitics of russia's annexation of crimea: narratives, identity, silences, and energy. *Eurasian Geography and Economics*, 55(3): 247–269, 2014. 1.5

[42] Susan Blackmore. *The meme machine*, volume 25. Oxford Paperbacks, 2000. 5.2.1

[43] Susan Blackmore, Lee Alan Dugatkin, Robert Boyd, Peter J Richerson, and Henry Plotkin. The power of memes. *Scientific American*, 283(4):64–73, 2000. 5.1

[44] Stephen Blank. Can information warfare be deterred? *Defense Analysis*, 17(2):121–138, 2001. 1.2.2

[45] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. 3.8.1, 4.2.1, 7.5.3

[46] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 3.2.6, 3.8.1, 7.3, 7.5.3, 7.5.3

[47] Liron Hakim Bobrov. Top 100 us media publications ranking h1 2018 by similarweb. `https://www.similarweb.com/blog/us-media-publications-ranking-h1-2018`, July 2018. (Accessed on 10/23/2019). 7.4.3

[48] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD*

International Conference on Knowledge Discovery & Data Mining, pages 71–79. ACM, 2018. 5.2.4

[49] Nellie Bowles. The mainstreaming of political memes online. *New York Times*, Feb 2018. URL `https://www.nytimes.com/interactive/2018/02/09/technology/political-memes-go-mainstream.html`. 5.1, A.4.6

[50] John Boyd. The d-n-i echo: The essence of winning and losing, by john r. boyd. `https://web.archive.org/web/20150223053831/http://www.danford.net/boyd/essence.htm`, 1995. (Accessed on 11/01/2019). 7.2

[51] Strategy Bridge. An extended discussion on an important question: What is information operations? `https://thestrategybridge.org/the-bridge/2017/5/8/an-extended-discussion-on-an-important-question-what-is-information-operations`. (Accessed on 12/28/2019). 1.2.2

[52] Ronald S Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009. 3.2.6

[53] Doyle Canning, Patrick Reinsborough, and Jonathan Matthew Smucker. *Re: Imagining change: How to use story-based strategy to win campaigns, build movements, and change the world*. Pm Press, 2017. 5.1

[54] Kathleen M. Carley. Group stability: A socio-cognitive approach. *Advances in Group Processes*, 7(1):44, 1990. B.2.2

[55] Kathleen M Carley, Michael K Martin, and Brian R Hirshman. The etiology of social change. *Topics in Cognitive Science*, 1(4):621–650, 2009. B.2.2

[56] Kathleen M Carley, Guido Cervone, Nitin Agarwal, and Huan Liu. Social cyber-security. In Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson, editors, *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018. 1.1

[57] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277, 1956. 3.2.6

[58] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018. 4.2.1

[59] Pablo Chamoso, Alberto Rivas, Javier J Martín-Limorti, and Sara Rodríguez. A hash based image matching algorithm for social networks. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 183–190. Springer, 2017. 5.4.2

[60] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. In *ICDM*, pages 817–822, 2016. 3.2, 3.2.2, 3.2.3, 4.1, 4.2.4, 6.2, A.4.5, B.1

[61] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. On-demand bot detection and archival system. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 183–187. International World Wide Web Conferences Steering

Committee, 2017. 3.3.4, 3.3

[62] Sergey G Chekinov and Sergey A Bogdanov. The nature and content of a new-generation war. *Military Thought*, 4:12–23, 2013. 1, 1.7

[63] Adrian Chen. The agency. *The New York Times*, 2(6):2015, 2015. A.2.1

[64] Chia-Mei Chen, DJ Guan, and Qun-Kai Su. Feature set identification for detecting suspicious urls using bayesian classification in social networks. *Information Sciences*, 289: 133–147, 2014. 3.2, 3.2.2

[65] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0133-6. doi: 10.1145/1920261.1920265. URL `http://doi.acm.org/10.1145/1920261.1920265`. 6.1

[66] Ben Cohen, Georgia Wells, and Tom McGinty. How one tweet turned pro-china trolls against the nba - wsj. `https://www.wsj.com/articles/how-one-tweet-turned-pro-china-trolls-against-the-nba-11571238943`, October 2019. (Accessed on 11/01/2019). 7.5.1

[67] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 5.2.2

[68] Nicholas Confessore, Gabriel JX Dance, Richard Harris, and Mark Hansen. The follower factory. *The New York Times*, 27, 2018. 6.2

[69] Michele Coscia. Competition and success in the meme pool: A case study on quickmeme. com. In *ICWSM*, 2013. 5.2.3

[70] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71, 2015. 3.3.4, 3.3

[71] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4): 561–576, 2018. 3.3.4, 3.3

[72] Stanley B Cunningham. *The idea of propaganda: A reconstruction*. Greenwood Publishing Group, 2002. 1.2.1

[73] Daryl J Daley and David G Kendall. Stochastic rumours. *Institute of Mathematics and Its Applications Journal of Applied Mathematics*, 1(1):42–55, 1965. B.2

[74] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016. 3.2.2, 4.6, 4.7, 6.2, A.4.5

[75] Nicola Davis. *The Selfish Gene*. Macat Library, 2017. 5.2.1

[76] Patrick Davison. The language of internet memes. *The social media reader*, pages 120–134, 2012. 5.1, 5.2.1

[77] Richard Dawkins. The selfish gene: with a new introduction by the author. *UK: Oxford University Press.(Originally published in 1976)*, 2006. 5.1, A.4.6

[78] Astolphe Louis Léonard de Custine and Pierre Nora. *Lettres de Russie: la Russie en 1839*. Gallimard, 1975. 1.3

[79] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific Reports*, 3:2980, 2013. B.2

[80] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. 4.2.1

[81] DJ Dekker. Measures of simmelian tie strength, simmelian brokerage, and, the simmelianly brokered. 2006. 3.2.6, 3.4.1

[82] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5.2.2

[83] Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. *The Tactics & Tropes of the Internet Research Agency*. New Knowledge, 2018. 4.7, 6.2, A.2.1

[84] J Donovan and B Friedberg. Source hacking: Media manipulation in practice. *Retrieved from Data&Society website: https://datasociety. net/output/source-hacking-media-manipulation-in-practice*, 2019. 5.2.1

[85] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. Memesequencer: Sparse matching for embedding image macros. In *Proceedings of the 2018 World Wide Web Conference*, pages 1225–1235. International World Wide Web Conferences Steering Committee, 2018. 5.2.2, 5.2.4, 5.4.4

[86] Robin IM Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(4):681–694, 1993. 3.3.2

[87] Phil Edwards. The reason every meme uses that one font - vox. `https://www.vox.com/2015/7/26/9036993/meme-font-impact`. (Accessed on 02/20/2019). 1

[88] T.S. Eliot. *The Rock*. Faber and Faber, London, 1934. 1.2

[89] Alessandro Epasto and Bryan Perozzi. Is a single embedding enough? learning node representations that capture multiple social contexts. In *The World Wide Web Conference*, pages 394–404. Acm, 2019. 4.2.2

[90] Robert Faris and Nart Villeneuve. Measuring global internet filtering. *Access denied: The practice and policy of global Internet filtering*, 5, 2008. A.2.2

[91] Russian Federation. Conceptual views regarding the activities of the armed forces of the russian federation in information space, 2016. 1.2.2, 1.3

[92] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8), 2017. 6.2

[93] Emilio Ferrara. Measuring social spam and the effect of bots on information diffusion in social media. *arXiv preprint arXiv:1708.08134*, 2017. 3.2.2, 3.2.4

[94] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 548–555. IEEE, 2013. 5.2.3

[95] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016. B.1

[96] Carol Taylor Fitz-Gibbon. *Performance indicators*, volume 2. Multilingual Matters, 1990. 7.4.3

[97] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010. 3.2.6

[98] Ove Frank. Sampling and estimation in large social networks. *Social networks*, 1(1): 91–101, 1978. 3.2.6

[99] David Mandell Freeman. Using naive bayes to detect spammy names in social networks. In *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*, pages 3–12. ACM, 2013. C.1

[100] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978. 3.2.6

[101] Linton C Freeman. Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, 3(3):291–304, 1982. 3.2.6

[102] Noah E Friedkin. *A Structural Theory of Social Influence*, volume 13. Cambridge University Press, 2006. B.2.2

[103] Kahina Gani, Hakim Hacid, and Ryan Skraba. Towards multiple identity detection in social networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 503–504. ACM, 2012. 3.2, 3.2.2

[104] James A Gavrilis. A model for population-centered warfare: a conceptual framework for analyzing and understanding the theory and practice of insurgency and counterinsurgency. *Small Wars Journal*, 10, 2009. 1.5

[105] Adam Geitgey. Face recognition. `https://github.com/ageitgey/face_recognition`, 2019. 5.3.1, 5.4.3

[106] Valery Gerasimov. The value of science is in the foresight. *Military Review*, 96(1):23, 2016. 1.1, 1.3

[107] Keir Giles. Handbook of russian information warfare. 2016. 1.2.2, 1.3

[108] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Practical recommendations on crawling online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892, 2011. 3.2.6

[109] April Glaser. Russian bots are trying to sow discord on twitter after charlottesville. 2017. 3.1

[110] Roy Gobson. Disinformation: A primer in russian active measures and influence campaigns. Written Testimony to the Senate Select Committee on Intelligence, Open Hearing, March 30, 2017, March 2017. 1.3

[111] Leo A Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961. 3.2.7, 4.6, 2, 7.3

[112] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018. 4.2.2

[113] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977. 3.2.6

[114] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social bots: Human-like by means of human control? *Big data*, 5(4):279–293, 2017. 3.1

[115] Elias Groll. How russia hacked u.s. politics with instagram marketing – foreign policy. `https://foreignpolicy.com/2018/12/17/how-russia-hacked-us-politics-with-instagram-marketing/`, December 2018. (Accessed on 04/06/2019). 5.5

[116] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 855–864. Acm, 2016. 4.2.2

[117] Alexus G Grynkewich. Introducing information as a joint function. Technical report, US Air Force Washington United States, 2018. 1.1, 1.2.2

[118] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008. 3.4, B.3

[119] Zellig S Harris. Methods in structural linguistics. 1951. 4.2

[120] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5.3.1

[121] Department of the Army Headquarters. The conduct of information operations. Technical report, Washington DC, 2018. 7.2

[122] *Offense and Defense: ADP 3-90*. Headquarters, Department of the Army, Washington DC, 2019 edition. 2

[123] Fritz Heider. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112, 1946. 3.2.6

[124] Todd C Helmus, Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega, and Zev Winkelman. *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*. Rand Corporation, 2018. 6.2

[125] Herbert W. Hethcote and James A. Yorke. *Gonorrhea Transmission Dynamics and Control*, volume 56. Springer, 1984. B.2

[126] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. 5.2.2

[127] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5.3.1

[128] Paul W Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 2(2):107–124, 1971. 3.2.6

[129] Paul W Holland and Samuel Leinhardt. A method for detecting structure in sociometric data. In *Social Networks*, pages 411–432. Elsevier, 1977. 3.2.6

[130] Philip N Howard and Bence Kollanyi. Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. *Browser Download This Paper*, 2016. 3.1, 6.2

[131] Philip N Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. *The IRA, social media and political polarization in the United States, 2012-2018*. University of Oxford, 2018. A.2.1

[132] Binxuan Huang and Kathleen M Carley. On predicting geolocation of tweets using convolutional neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 281–291. Springer, 2017. 3.8

[133] Cherilyn Ireton and Julie Posetti. *Journalism, fake news & disinformation: handbook for journalism education and training*. UNESCO Publishing, 2018. 6.1

[134] Sylvio Barbon Jr, Gabriel F. C. Campos, Gabriel M. Tavares, Rodrigo A. Igawa, Mario L. Proença Jr, and Rodrigo Capobianco Guido. Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. *ACM Trans. Multimedia Comput. Commun. Appl.*, 14(1s):26:1–26:17, March 2018. ISSN 1551-6857. doi: 10.1145/3183506. URL `http://doi.acm.org/10.1145/3183506`. 6.1

[135] Robert D Kaplan. The revenge of geography. *Foreign Policy*, (172):96–105, 2009. 1.5

[136] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1): 39–43, 1953. 3.2.6

[137] Franziska Keller, David Schoch, Sebastian Stier, and JungHwan Yang. Political astroturfing on twitter: How to coordinate a disinformation campaign. 2019. 6.2

[138] M Khayat, M Karimzadeh, J Zhao, and DS Ebert. Vassl: A visual analytics toolkit for social spambot labeling. *IEEE transactions on visualization and computer graphics*, 2019. 6.2

[139] Carolyn Mae Kim. *Social media campaigns: Strategies for public relations and marketing*. Routledge, 2016. 7.2

[140] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolu-

tional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4.2.3

[141] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302, 2015. 4.2.1

[142] David Krackhardt. The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations*, 16(1):183–210, 1999. 3.2.6, 3.4.1

[143] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5.2.2

[144] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *arXiv preprint arXiv:1802.04289*, 2018. 3.1, 3.2, 3.2.2

[145] UE Kuleshov, BB ZHUTDIEV, and YES Fedorov. Information-psychological confrontation in modern conditions: theory and practice. *Bulletin of the Academy of Military Sciences*, (1):104–110, 2014. 1.3

[146] Sumeet Kumar and Kathleen Carley. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5047–5058, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1498. URL `https://www.aclweb.org/anthology/P19-1498`. 5, 7.3

[147] Sumeet Kumar and Kathleen M Carley. What to track on the twitter streaming api? a knapsack bandits approach to dynamically update the search terms. 2019. URL `https://asonamdata.com/ASONAM2019_Proceedings/pdf/papers/023_0158_070.pdf`. 1, 7.3

[148] Harold D Lasswell. The strategy of soviet propaganda. *Proceedings of the Academy of Political Science*, 24(2):66–78, 1951. 1.3, 1.5

[149] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010. 3.2, 3.2.2

[150] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*, 2011. 3.3.4

[151] Sangho Lee and Jong Kim. Early filtering of ephemeral malicious accounts on twitter. *Computer Communications*, 54:48–57, 2014. 3.2, 3.2.2

[152] Vladimir Lefebvre. *Conflicting structures*. Lulu. com, 2015. 1.3

[153] Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. Pytorch-biggraph: A large-scale graph embedding system. *arXiv preprint arXiv:1903.12287*, 2019. 4.2.2

[154] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009. 5.2.3

[155] Mei Li, Xiang Wang, Kai Gao, and Shanshan Zhang. A survey on information diffusion in online social networks: Models and methods. *Information*, 8(4):118, 2017. B.2

[156] Andrew Lih. *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. Hachette Books, 2009. 1.2

[157] Dehai Liu, Benjin Mei, Jinchuan Chen, Zhiwu Lu, and Xiaoyong Du. Community based spammer detection in social networks. In *International Conference on Web-Age Information Management*, pages 554–558. Springer, 2015. 3.2.4

[158] Ting Liu, Charles Rosenberg, and Henry A Rowley. Clustering billions of images with large scale nearest neighbor search. In *2007 IEEE Workshop on Applications of Computer Vision (WACV'07)*, pages 28–28. IEEE, 2007. 5.4.2

[159] Al Bawaba The Loop. Thousands of twitter bots are attempting to silence reporting on yemen. 2017. 2.3, 3.1, 4.3.2, C.1.1

[160] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 5.4.2

[161] Cristian Lumezanu, Nick Feamster, and Hans Klein. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012. 2.3, 3.1

[162] Thomas Magelinski, David Beskow, and Kathleen M. Carley. Graph-hist: Graph classification from latent feature histograms with application to bot detection. *Association for the Advancement of Artificial Intelligence*, 2020. 3.9, 4.2.4

[163] Jay Mahadeokar, Sachin Farfade, Amar Ramesh Kamat, and Armin Kappeler. 6.5.1

[164] Daniel P Maki and Maynard Thompson. Mathematical models and applications: With emphasis on the social life, and management sciences. Technical report, 1973. B.2

[165] William Marcellino, Meagan L Smith, Christopher Paul, and Lauren Skrabala. *Monitoring Social Media*. RAND Corporation, 2017. 1.7

[166] Peter Mattis. Contrasting china's and russia's influence operations. *War on the Rocks*, 2018. 1.4

[167] Kevin N McCauley. *Russian Influence Campaigns Against the West: From the Cold War to Putin*. CreateSpace Independent Publishing Platform, 2016. 1.3

[168] Stephen McDonell. Why china censors banned winnie the pooh - bbc news. `https://www.bbc.com/news/blogs-china-blog-40627855`, July 2017. (Accessed on 04/06/2019). 1.4, 5.5

[169] Stephen McDonell. Why china censors banned winnie the pooh - bbc news. `https://www.bbc.com/news/blogs-china-blog-40627855`, July 2017. (Accessed on 09/29/2019). A.4.6

[170] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. B.3.3

[171] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 4.2.1

[172] Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73, 2014. 3.2.2

[173] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. Botwalk: Efficient adaptive exploration of twitter bot networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 467–474. ACM, 2017. 6.2

[174] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007. 3.2.6

[175] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *Proceedings of the Seventh International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, Palo Alto, California, 2013. Association for the Advancement of Artificial Intelligence Press. B.5

[176] Robert S Mueller. Report on the investigation into russian interference in the 2016 presidential election. *US Dept. of Justice. Washington, DC*, 2019. 4.7, A.2.1

[177] Dieudonne Mulamba, Indrajit Ray, and Indrakshi Ray. Sybilradar: A graph-structure based framework for sybil detection in on-line social networks. In *IFIP International Information Security and Privacy Conference*, pages 179–193. Springer, 2016. 3.11

[178] Mahdi Namazifar. Detecting randomly generated strings, December 2015. [Online: posted 25 December 2015]. C.1

[179] Engineering National Academies of Sciences and Medicine. *A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis*. The National Academies Press, Washington, DC, 2019. ISBN 978-0-309-48761-0. doi: 10.17226/25335. URL `https://www.nap.edu/catalog/25335/a-decadal-survey-of-the-social-and-behavioral-sciences-a`. A.1

[180] LM Neudert, B Kollanyi, and PN Howard. Junk news and bots during the german federal presidency election: What were german voters sharing over twitter?, 2017. 3.1

[181] Ben Nimmo. Anatomy of an info-war: How russia's propaganda machine works, and how to counter it. *Central European Policy Institute*, 15, 2015. 2

[182] Ben Nimmo. #botspot: Twelve ways to spot a bot – dfrlab – medium. `https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c`, 11 2017. (Accessed on 11/03/2018). 3.4.3, 6.1, 6.2, 6.5.1

[183] Ben Nimmo, Graham Brookie, and Kanishk Karan. #trolltracker: Twitter troll farm archives - dfrlab - medium. `https://medium.com/dfrlab/trolltracker-twitter-troll-farm-archives-8d5dd61c486b`. (accessed on 09/23/2019). A.2.1

[184] Carol L Novak and Steven A Shafer. Anatomy of a color histogram. In *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 599–605. IEEE, 1992. 5.4.2

[185] Department of Defense. Strategy for operations in the information environment. Technical report, Washington DC, 2016. 7.1

[186] Stephen M Omohundro. *Five Balltree Construction Algorithms*. International Computer Science Institute Berkeley, 1989. 5.4.2

[187] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 1105–1114. Acm, 2016. 4.2.2

[188] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*, 2017. 4.2.1

[189] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 3.5, C.2.1

[190] V Peirson, L Abel, and E Meltem Tolunay. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*, 2018. 5.1, 5.2.2

[191] James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011. 6.5.2

[192] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (emnlp)*, pages 1532–1543, 2014. 4.2.1, 5.3.1

[193] Nathaniel Persily. The 2016 us election: Can democracy survive the internet? *Journal of democracy*, 28(2):63–76, 2017. 6.2

[194] Ben Popken. Russian trolls duped global media and nearly 40 celebrities. `https://www.nbcnews.com/tech/social-media/trump-other-politicians-celebs-shared-boosted-russian-troll-tweets-n817036`, November 2017. (Accessed on 10/23/2019). 7.4.3

[195] Jarred Prier. Commanding the trend: Social media as information warfare. *Strategic Studies Quarterly*, 11(4), 2017. 1.2.2

[196] Jayaram Raghuram, David J Miller, and George Kesidis. Unsupervised, low latency anomaly detection of algorithmically generated domain names by generative probabilistic modeling. *Journal of advanced research*, 5(4):423–433, 2014. C.1

[197] Clifford Reid. Reflexive control in soviet military planning. *Soviet Strategic Deception*, pages 293–312, 1987. 1.3

[198] HAN Rongbin. Manufacturing consent in cyberspace: China's "fifty-cent army". *Journal of Current Chinese Affairs*, 44(2):105–134, 2015. 1.2.2, 1.4

[199] Jennifer Rowley. The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science*, 33(2):163–180, 2007. 1.2

[200] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019. 4.2.1

[201] Fabio Rugge. ”mind hacking”: Information warfare in the cyber age — ispi. *Analysis*, (319), January 2009. 1.6

[202] Ronald Rychiak and Ion Mihai Pacepa. *Disinformation: Former Spy Chief Reveals Secret Strategies for Undermining Freedom, Attacking Religion, and Promoting Terrorism*. Washington, Dc: WND books, 2013. 1.3

[203] Twitter Safety. Information operations directed at hong kong. `https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html`, August 2019. (Accessed on 09/26/2019). 7.5.1, A.2.2

[204] Badrul Munir Sarwar, George Karypis, Joseph A Konstan, John Riedl, et al. Item-based collaborative filtering recommendation algorithms. *WWW*, 1:285–295, 2001. 4.2.1

[205] Kevin D Scott. Joint operating environment 2035: The joint force in a contested and disordered world. Technical report, Joint Chiefs of Staff Washington United States, 2016. 1.2.2

[206] Emilio Serrano, Carlos A Iglesias, and Mercedes Garijo. A survey of twitter rumor spreading simulations. In *Computational Collective Intelligence*, pages 113–122. Springer, 2015. B.2.1

[207] Emilio Serrano, Carlos Ángel Iglesias, and Mercedes Garijo. A novel agent-based rumor spreading model in twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW ’15 Companion, pages 811–814, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2742466. URL `http://doi.acm.org/10.1145/2740908.2742466`. B.2.1, B.3

[208] Scott Shane and Sheera Frenkel. Operation targeted african-americans on social media. 2018. 4.3.3

[209] Claude E Shannon. The bell system technical journal. *A mathematical theory of communication*, 27:379–423, 1948. C.1

[210] Claude E Shannon and WARREN Weaver. The mathematics theory of communication. *Bell Syst. Tech. J*, 27:379–423, 1948. 1.2

[211] Elisa Shearer and Jeffrey Gottfried. News use across social media platforms 2017. *Pew Research Center, Journalism and Media*, 2017. 1.6.1

[212] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011. 4.2.2

[213] Limor Shifman. An anatomy of a youtube meme. *new media & society*, 14(2):187–203, 2012. 5.1

[214] Limor Shifman. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, 18(3):362–377, 2013. 5.2.1

[215] Limor Shifman. The cultural logic of photo-based meme genres. *Journal of Visual Culture*, 13(3):340–358, 2014. 5.2.1, A.4.6

[216] Limor Shifman. *Memes in digital culture*. MIT press, 2014. 5.2.1, A.4.6

[217] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5.3.1

[218] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 5.3.3

[219] Ray Smith. An overview of the tesseract ocr engine. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 629–633. IEEE, 2007. 5.3.1

[220] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. Neural compatibility modeling with attentive knowledge distillation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 5–14. ACM, 2018. 5.2.2

[221] JCo Staff. Information operations joint publication 3-13. *Washington, DC: Department of Defense*, 2012. 1.2.2

[222] Joint Staff. Information operations. Technical report, Washington DC, 2012. (document), 7.2, 7.1

[223] Joint Staff. Doctrine for the armed forces of the united states (jp 1). Technical report, US Department of Defense, 2017. 1.2.2

[224] Jack Stubbs and Johan Ahlander. Exclusive: Right-wing sites swamp sweden with 'junk news' in tight election race — reuters. `https://www.reuters.com/article/us-sweden-election-disinformation-exclus/exclusive-right-wing-sites-swamp-sweden-with-junk-news-in-tight-election-race-idUSKCN1LM0DN`, 09 2018. (Accessed on 11/20/2018). 3.8

[225] Feng Su and Like Xue. Graph learning on k nearest neighbours for automatic image annotation. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 403–410. ACM, 2015. 5.4.2

[226] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016. 3.2.2

[227] Marcella Szablewicz. The 'losers' of china's internet: Memes as 'structures of feeling'for disillusioned young netizens. *China Information*, 28(2):259–275, 2014. 5.1

[228] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 5.3.1

[229] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 earthquake: a twitter case study. *PLoS one*, 10(4):e0121443, 2015. B.2

[230] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM, 2011. 3.3.4

[231] Elizabeth Thompson. Most of canada's top websites won't post federal election ads this year — cbc news. `https://www.cbc.ca/news/politics/online-election-advertising-canada-1.5116753`, May 2019. (Accessed on 11/29/2019). 4.7

[232] Stephen Townsend. Accelerating multi-domain operations: Evolution of an idea - modern war institute. `https://mwi.usma.edu/accelerating-multi-domain-operations-evolution-idea/`, July 2018. (Accessed on 08/16/2018). 1.1

[233] Eleni Triantafillou, Jamie Ryan Kiros, Raquel Urtasun, and Richard Zemel. Towards generalizable sentence embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 239–248, 2016. 4.2.1

[234] Rudra M. Tripathy, Amitabha Bagchi, and Sameep Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Conference of Information and Knowledge Mangement '10, pages 1817–1820, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871737. URL `http://doi.acm.org/10.1145/1871437.1871737`. B.2, B.2.1

[235] William Thomas Tutte. Convex representations of graphs. *Proceedings of the London Mathematical Society*, 3(1):304–320, 1960. 4.2

[236] Twitter. Elections integrity data archive. `https://about.twitter.com/en_us/values/elections-integrity.html#us-elections`. (Accessed on 03/30/2019). 3.3.4

[237] Twitter. Twitter election integrity data archive. `https://about.twitter.com/en_us/values/elections-integrity.html#data`, 2018. Accessed: 2018-11-30. 4.3.3

[238] Tom Uren, Elise Thomas, and Jacob Wallis. *Tweeting Through the Great Firewall: Preliminary Analysis of PRC-linked Information Operations on the Hong Kong Portest*. Australia Strategic Policy Institute: Internationl Cyber Policy Center, 2019. 6.2, A.2.2

[239] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*, 2017. 3.2, 3.3.4, 3.5

[240] John-Paul Verkamp and Minaxi Gupta. Five incidents, one theme: Twitter spam as a weapon to drown voices of protest. In *FOCI*, 2013. 2.3, 3.1

[241] Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. An analysis of social network-based sybil defenses. *ACM SIGCOMM Computer Communication Review*,

40(4):363–374, 2010. 3.11

[242] Bimal Viswanath, Muhammad Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Towards detecting anomalous user behavior in online social networks. In *USENIX Security Symposium*, pages 223–238, 2014. 3.2.2

[243] Gerret Von Nordheim, Karin Boczek, and Lars Koppers. Sourcing the sources: An analysis of the use of twitter and facebook as a journalistic source over 10 years in the new york times, the guardian, and süddeutsche zeitung. *Digital Journalism*, 6(7):807–828, 2018. A.3

[244] Edward Waltz. *Information warfare: Principles and operations*. Artech House Boston, 1998. (document), 7.2, 7.2

[245] Chao Wang, Zong Xuan Tan, Ye Ye, Lu Wang, Kang Hao Cheong, and Neng-gang Xie. A rumor spreading model based on information entropy. *Scientific reports*, 7(1):9615, 2017. B.2.1, B.3.4

[246] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234. Acm, 2016. 4.2.2

[247] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*, 2012. 3.2, 3.2.3

[248] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014. 5.4.2

[249] William Yang Wang and Miaomiao Wen. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–365, 2015. 5.2.2

[250] Clint Watts. *Messing with the enemy: Surviving in a social media world of hackers, terrorists, Russians, and fake news*. Harper Business, 2018. 1.6.1, 6.5.1

[251] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2:335, 2012. B.3.2

[252] Norbert Wiener. Cybernetics or control and communication. *Bell System Tech. J*, 27, 1948. 1.2

[253] Jasmyn Wimbish and Colin Ward-Henninger. Nba-china issue: Latest news resulting from daryl morey's hong kong tweet, what it means for the league - cbssports.com. `https://www.cbssports.com/nba/news/nba-china-issue-latest-news-resulting-from-daryl-moreys-hong-kong-tweet-what-it-`

means-for-the-league/, October 2019. (Accessed on 11/01/2019). 7.5.1

[254] Stefan Wojcik, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. Bots in the twittersphere. *Pew Research Center. Retrieved May*, 22:2018, 2018. 6.2

[255] Ling-Ling Xia, Guo-Ping Jiang, Bo Song, and Yu-Rong Song. Rumor spreading model considering hesitating mechanism in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 437:295–303, 2015. B.2

[256] Cao Xiao, David Mandell Freeman, and Theodore Hwa. Detecting clusters of fake accounts in online social networks. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, pages 91–101. ACM, 2015. 4.2.4

[257] Lexing Xie, Apostol Natsev, John R. Kender, Matthew Hill, and John R. Smith. Visual memes in social media: Tracking real-world news in youtube videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 53–62, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0616-4. doi: 10.1145/2072298.2072307. URL http://doi.acm.org/10.1145/2072298.2072307. 5.2.2

[258] Sandeep Yadav, Ashwath Kumar Krishna Reddy, AL Reddy, and Supranamaya Ranjan. Detecting algorithmically generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 48–61. ACM, 2010. C.1

[259] Joshua Yaffa. Dmitry kiselev, putin's favorite tv host, is russia's top propogandist — the new republic. https://newrepublic.com/article/118438/dmitry-kiselev-putins-favorite-tv-host-russias-top-propogandist, July 2014. (Accessed on 08/16/2018). 1.1

[260] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019. 3.10

[261] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. *arXiv preprint arXiv:1911.09179*, 2019. 4.6

[262] Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 267–278. ACM, 2006. 3.2.4, 3.11, 6.2

[263] Damian H Zanette. Dynamics of rumor propagation on small-world networks. *Physical Review E*, 65(4), 2002. B.2

[264] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202. ACM, 2018. 5.2.3, 5.2.4, 5.4.4

[265] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out?: Towards understanding state-sponsored trolls. In *Proceedings of the 10th ACM Conference on Web Science*, pages 353–362. ACM, 2019. 6.2

[266] Peter Zeihan. *The accidental superpower: the next generation of American preeminence and the coming global disorder*. Twelve, 2014. 1.5

[267] Laijun Zhao, Qin Wang, Jingjing Cheng, Yucheng Chen, Jiajia Wang, and Wei Huang. Rumor spreading model with consideration of forgetting mechanism: A case of online blogging livejournal. *Physica A: Statistical Mechanics and its Applications*, 390(13): 2619–2625, 2011. B.2

[268] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651. IEEE, 2017. 5.3.2

[269] Chaim Zins. Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology*, 58(4):479–493, 2007. 1.2