Institute of Information Security

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

# Disinformation Campaigns in Social Media

Robin Sliwa

| | |
|---|---|
| **Course of Study:** | Informatik |
| **Examiner:** | Prof. Dr. Ralf Küsters |
| **Supervisor:** | Dipl.-Inf. Guido Schmitz |
| **Commenced:** | October 23, 2019 |
| **Completed:** | June 29, 2020 |

## Abstract

In an increasingly digitally connected world, social networks have become a large factor in news consumption, discussion and staying connected to friends. This thesis aims to give an overview over how this new platform has been a vector for the conduction of disinformation campaigns. Beyond the prime example - possible Russian disinformation in the U.S. from 2015 to 2017 - and its efficacy, further candidates as well as the historical context, technical aspects and the public response are touched upon. The U.S. election of 2016 is evidently a well-documented example of an election targeted by a large-scale disinformation campaign conducted through social media. Indications exist that campaigns are also being conducted in other political contexts (France, 2017) and with contexts extending into economics. This thesis also finds that more research is needed to systematically detect and investigate disinformation campaigns, especially in order to measure and contain their real-world impact.

# Contents


**1 Introduction** **11**

**2 Disinformation Campaigns** **13**
2.1 Historic Development . . . 13
2.1.1 Operation Denver . . . 13
2.1.2 Western Disinformation . . . 15
2.1.3 Online Information Campaigns . . . 16
2.2 The US Election of 2016 . . . 18
2.2.1 Timeline of Discovery . . . 19
2.2.2 The IRA . . . 20
2.2.3 Facebook Operations . . . 21
2.2.4 Instagram Operations . . . 25
2.2.5 Twitter Operations . . . 27
2.2.6 Activity Drop of 2017 and After . . . 29
2.3 Efficacy of IRA activity . . . 30
2.4 Impact in the European Union . . . 33
2.4.1 2017 French Presidential Election . . . 33
2.4.2 Brexit . . . 35
2.5 Big Data . . . 36
2.5.1 Cambridge Analytica . . . 37
2.5.2 Obama for America App . . . 39
2.6 Beyond Political Campaigns . . . 40

**3 Technical Aspects** **43**
3.1 Anonymization . . . 43
3.1.1 VPN . . . 43
3.1.2 Proxies . . . 44
3.1.3 Tor . . . 45
3.2 Deanonymization . . . 47
3.2.1 Super Cookies . . . 48
3.2.2 Fingerprinting . . . 48
3.2.3 Traffic Analysis . . . 50
3.3 Deepfakes . . . 51
3.3.1 Faceswaps . . . 51
3.3.2 Synthesized Speech . . . 52
3.4 Social Bots . . . 52
3.5 Countermeasures by Social Networks . . . 54
3.5.1 Twitter . . . 54
3.5.2 Facebook . . . 55

3.6 Automated Detection of Disinformation . . . . . 56

**4 Public Response** **59**
4.1 Australia . . . . . 59
4.2 California Consumer Privacy Act (CCPA) . . . . . 59
4.3 General Data Protection Regulation (GDPR) . . . . . 62
4.4 Germany . . . . . 63
4.4.1 Network Enforcement Act (NetzDG) . . . . . 63
4.4.2 Tagesschau Faktenfinder . . . . . 64

**5 Conclusion and Outlook** **65**

**Bibliography** **67**

# List of Figures


2.1 Usage of the term disinformation in English literature published between 1900 and 2000, as measured and plotted as percentage of all words in scanned English books by Google Books Ngram Viewer . . . . . 14
2.2 Excerpt from the report "Behavioural Science Support for JTRIG's (Joint Threat Research and Intelligence Group's) Effects and Online HUMINT Operations", leaked by Edward Snowden and published by The Intercept [Dha11]. . . . . 17
2.3 Example: The Army of Jesus group on Facebook was created by Internet Research Agency (IRA) operatives, with the IRA themselves purchasing advertisements on Facebook to reach more users during the U.S. election cycle of 2016. Source: Facebook advertisements released by United States House Permanent Select Committee on Intelligence (HPSCI) . . . . . 22
2.4 Amount of IRA advertisements on Facebook, by month and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.15] . . . . . 24
2.5 Amount of IRA advertisements on Facebook, by day and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.15] . . . . . 24
2.6 Amount of IRA posts on Facebook, by day and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.16] . . . . . 25
2.7 Amount of IRA posts on Instagram, by day and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.16] . . . . . 26
2.8 Amount of IRA tweets on Twitter targeted at Russia (orange) and the United States (blue) between 2009 and 2018. Image source: [HGL+19, p.28] . . . . . 27
2.9 IRA Tweets on Twitter with different types of focus - regarding local events or political ideologies between 2013 and 2018. Image source: [HGL+19, p.28] . . . 28
2.10 IRA Number of tweets per hour starting from October 5th, 2016, 12:00 AM UTC. Number of tweets by accounts categorized as politically left (grey) / right (black) trolls is plotted, with high activity appearing before the release of John Podesta's emails by Wikileaks. Image source: [BLW18, p.19] . . . . . 29
2.11 Monthly average amount of IRA posts on Twitter, Facebook and Instagram by year, from 2015 to 2017. Data source: [HGL+19, p.5] . . . . . 30
2.12 The measured effect of following the opposing-views bot on Twitter by the respondents. The dots represent the estimated averages, with the red (Republican) and blue (Democrats) lines denoting the 95% (bold) and 90% (slim) confidence intervals. Image source: [BAB+18, p. 9220] . . . . . 31
2.13 Results of interaction with IRA-related tweets over time - the points correspond to average effect of exposure, and the blue lines the 95% credible interval. Image source: [BGM+20, p. 246] . . . . . 32
2.14 Generic election (grey) and MacronLeaks (purple) related tweets during April 25th and May 7th, 2017. The tweets counted were related to the French election campaign and MacronLeaks. Time in UTC. Image source: [Fer17, p.29] . . . . . 34

2.15 Map displaying the voter shift in each US district in the 2016 U.S. election. Source: Larry Buchanan et al. for The New York Times, based on election results data compiled by the Associated Press, color-adjusted by author for better visibility . . 37
2.16 Development of Facebook's stock price between January and May 2018. The red line denotes day the Cambridge Analytica (CA) story broke on the NYT and The Guardian. Data via Yahoo! Finance . . . . . 38

3.1 Simplified illustration of VPN functionality, with two-sided arrows indicating data connections, and additional locks denoting the connection to be secured by encryption. 44
3.2 Illustration of Tor connection. The cloud and computers inside it illustrate the wider Tor network, with the red arrows denoting the current Tor circuit. Also present outside the network is the directory server and the actual connection destination. . 45
3.3 Illustration of the Agent, Message, Interpreter conceptual framework, with the smaller elements on the right denoting attributes and possible values. Image source: [WD17] . . . . . 57

# Acronyms

**AMI** Agent Message Interpreter. 57

**CA** Cambridge Analytica. 8

**CCPA** California Consumer Privacy Act of 2018. 60

**FBI** Federal Bureau of Investigation. 19

**GDPR** General Data Protection Regulation. 39

**HPSCI** United States House Permanent Select Committee on Intelligence. 7

**IRA** Internet Research Agency. 7

**SSCI** United States Senate Select Committee on Intelligence. 19

# 1 Introduction

Public opinion has been the key to holding power in democracies - and to an extent autocracies - for centuries. Both domestic and foreign powers have tried to leverage this by undermining opponents or improving their own image for their gain. The rise of the internet and social media specifically, however, have given way to entirely new methods of doing so. In social networks such as Facebook and Twitter, users can interact freely with one another and spread their content with the use of hashtags[1], which opens up new ways to insert wrong information - and reach the masses with it. Meanwhile, big data has lately enabled improved audience targeting, and artificial intelligence has undermined the weight of video and audio evidence by creating new ways to doctor records. This thesis aims to gain insight into the following questions:

1. What examples for social media disinformation campaigns are there?
2. How are these campaigns conducted?
3. Is there a way to combat them, and if so, how?
4. How have regulators responded?

This thesis is structured in the following way:

**Chapter 1 - Disinformation Campaigns** introduces disinformation from a historical context, followed by the centerpiece of this thesis, an analysis of the 2016 U.S. election, including a look at disinformation activity in different social networks. Also, the connection to Big Data for higher impact by improved audience targeting is made, and relevance in economic lobbying established also.

**Chapter 2 - Technical Aspects** describes techniques utilized for disinformation campaigns, divided in two categories: Technologies to better mask disinformation by anonymization as well as deepfakes for the creation of near-authentic false information as content to spread. Followed by that, responses by social media companies are described and methods for detection of campaigns established.

**Chapter 3 - Public Response** analyzes measures taken by governments and legislatures to combat disinformation, and additionally, related aspects such as Big Data uses for voter targeting.

**Chapter 4 - Conclusion and Outlook** returns to the questions raised, providing answers and summing up key knowledge gained in this thesis. Furthermore, an outlook and recommendations for further search are made as well.

[1] A hashtag is a phrase preceded by a hash sign, used on social media (especially Twitter) to link specific content to a topic.

# 2 Disinformation Campaigns

Disinformation is a type of false information, with the deliberate intent to use a lie in order to mislead the recipient of the information [Web19]. In the scope of this thesis, we will focus on the deliberate spread of false information rather than simple misinformation - which is similar, but sans the intent and awareness of falseness. Meanwhile, propaganda is the act of spreading any information - rumors, facts, ideas or lies - with the intent of supporting or damaging a cause, person or institution. It can - but does not have to - use false information for this purpose and often measures are taken to mask that it is propaganda (e.g. cloaking the source or boosting the seeming validity of the information). The borders between propaganda, misinformation and disinformation are, however, at times rather blurred. Still, various forms of information warfare have been present throughout human civilization, though the modern methods have changed drastically.

## 2.1 Historic Development

One of the earliest recorded examples occurred during the last days of the Roman Republic. There are known smear campaigns initiated by Octavian, later known as Emperor Augustus, versus his rival Marcus Antonius [Sco33]. Using an alleged will of Marcus Antonius, easily providing ample content to brand him as being under the influence of Egypt's Cleopatra, he convinced the Roman Senate and the people of Antonius being a traitor and thus bringing him the upper hand in the struggle for power. Whether it was genuine or at least partially forged is still disputed. Either way, this can be considered as one of the earliest documented instances of swaying public opinion in order to gain advantage over an opponent by constructing a negative image of them.

### 2.1.1 Operation Denver

Using false information to achieve advantages has, of course, been a factor in human society for a long time. The term "disinformation", however, was created in the 1900s in the Soviet Union and is said to be coined by Joseph Stalin. In fact, the term itself started getting traction in literature (e.g. in the Great Soviet Encyclopedia of 1952), as can be seen in Figure 2.1. It reached its peak usage in the 1980s, when it became a public topic in the US during high profile topics such as Operation Denver, which will be touched upon later, and the administration of Ronald Reagan's measures to counter Soviet efforts to spread false information (and use similar tactics offensively as well).

While those may count for the first documented attempts at information warfare and disinformation campaigns, a more recognizable, modern approach started to arise in the 20th century. As implied by the origin of the word disinformation in the Soviet Union, the superpower put a lot of effort into controlling the flow of information - domestically and internationally, through the KGB [Rom01]. Soviet defectors played a large role in helping the West understand covert operations during the

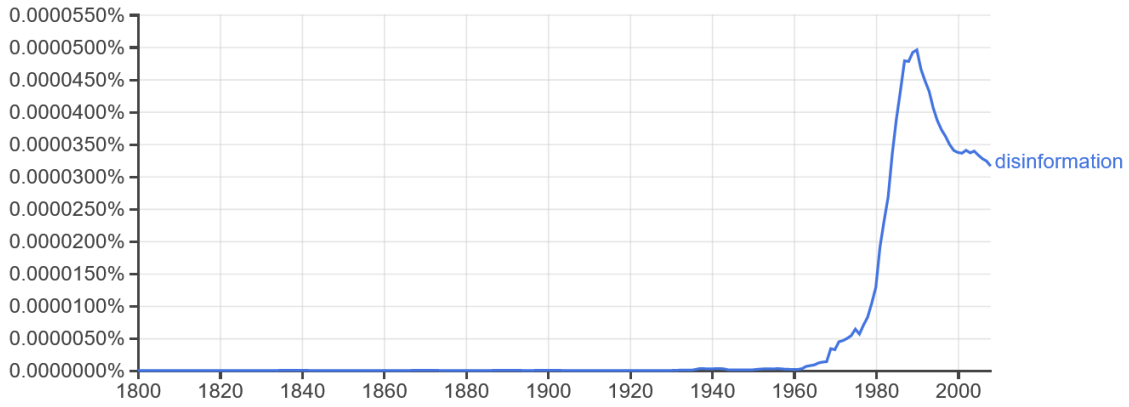


**Figure 2.1:** Usage of the term disinformation in English literature published between 1900 and 2000, as measured and plotted as percentage of all words in scanned English books by Google Books Ngram Viewer

Cold War. A well-documented attempt at doing so was Operation DENVER (popularly also known as INFEKTION) which aimed to influence world opinion towards the belief that HIV/AIDS had been created by the United States. The operation's impact commenced in July 1983, when an article was published in the Indian newspaper Patriot. The newspaper itself has been claimed to be created by the KBG in order to spread disinformation [Uni87]. In 1985, based on this groundwork, the KGB started a series of actions (active measures) to make the international public believe that AIDS had resulted from biological weapon experiments conducted by the US gone wrong, as stated in a telegram requesting assistance from their allied Bulgarian counterparts [Sta85].

Soon after, a Soviet newspaper called Literaturnaya Gazeta, known outlet for the KGB, published another article. The article by Valentin Zapevalov made allegations regarding the US having developed HIV. This time, not only the Pentagon was claimed to have been involved, but also the Center for Disease Control (CDC) was alleged to have helped with sample collection in Africa. Collected pathogenic samples had thus been used to create a combined virus that had been entirely foreign to Europe and Asia. US-conducted human tests in Haiti on marginalized and vulnerable groups such as drug-addicts and homosexuals were claimed to have happened as well. The article was later reprinted in multiple countries such as Kuwait, Bahrain, Finland, Sweden and Peru [Cen86]. However, this was only one step in the campaign.

From there on out, a version of the conspiracy theory was further publicized by the East German emerited biology professor Jakob Segal. With the seemingly scientific backing, the entire theory gained new momentum with interviews in Western press such as the British The Sunday Express. Allegations described the alleged HIV synthesis out of two pre-existing viruses perpetrated by the U.S., Whether or not Segal acted under guidance of the KGB or the Stasi is debated until this day, with inconclusive results. Regardless of that, the soviet disinformation campaign showed measurable results. The conspiracy theory was amplified by newspaper reports, magazine reports, radio coverage and television features in more than 30 languages and 80 countries. The Soviets specifically targeted African countries, in which they possessed wide influence with free content distribution as well as benefits for publicizing propaganda. [Uni87]

The campaign showed notable success: In 1992, 15% of US-Americans thought it was definitely or likely true that HIV had been created deliberately in a government laboratory. In 2005, a study showed that almost 50% of US-Americans thought AIDS was man-made, while over 25% considered it to have been created in a government laboratory - and 15% thought it was a form of genocide against black people [Bog09, p.19]. This well documented example of Cold War era disinformation showed a great potential for swaying public opinion, even before the level of international connection available since the rise of the internet.

### 2.1.2 Western Disinformation

It should be noted that during political campaigning, there is the potential for grey areas. Those may exist in between simple political campaigning and positioning and clear spread of false information. Especially in polarized situations and between parties with opposite views, especially when they are perceived as competitive with one another. There are non-neutral newspapers openly associated with and published by parties, such as the Bayernkurier in Germany. The Bayernkurier was the conservative Christian Social Union in Bavaria's (CSU) party newspaper from 1950 until 2019. In it, the party would publish articles around all sorts of political topics. Often, they would clash with the Social Democratic Party (SPD), Germany's traditional politically center-left party. Reporting about the SPD was usually negative and not neutral, and as such topics were selected specifically to give the CSU a profile as better option and to degrade the SPD in the perception of the readers - party members and affiliated people. For example, the association of the SPD with socialism was generally pursued and the taboo of cooperating with the Party of Democratic Socialism (PDS), successor to the East German ruling party. Ultimately, this practice might have been polarizing but did not generally cross the line towards disinformation [Wag00, p.129 - 133]. It is therefore important to carefully distinguish regular political discourse - even if it is polarizing - from purposeful and strategic disinformation campaigns.

#### Reagan Administration

It is important to note that disinformation tactics are not limited to Authoritarian countries, parties or politicians. There are examples of Western democracies engaging in those activities as well. A popular example in a similar timeframe as the aforementioned Operation Denver / INFEKTION would be disinformation by the administration of U.S. president Ronald Reagan [Bit90]. Specifically, in 1986, Libya and its dictator Gaddafi were the target of false information spread by U.S. officials and disseminated through the press. In particular, the White House and State Department were accused of having formalized and pursued a policy of spreading false information giving the impression that the U.S. would pursue an imminent attack or regime change in Libya [Gel86]. Relations between the West and Libya had been strained. The U.S. had been displeased ever since Libya nationalized its oil industry after the 1969 coup that put Gaddafi in power. Later, multiple terrorist attacks occurred in which there were ties to Libya, leading to embargos and entry to the list of "state sponsors of terrorism" by the U.S. Most recently, in January 1986, a discotheque in West Berlin was bombed, killing two U.S. servicemen, and the U.S. responded with air strikes [McC87]. In August 1986, both the Washington Post and The Wall Street Journal quoted anonymous officials intentionally and falsely saying that the U.S. and Libya were again on a collision course. This tactic was denounced

by the American Society of Newspaper Editors and the media[Gel86]. This is therefore a prime example that Western democracies are not the only targets of disinformation - they can employ the same tactics, crossing ethical lines doing so.

### 2.1.3 Online Information Campaigns

As the internet rose to a primary source of discourse and information retrieval, its user base expanded rapidly. This shift of public discourse gave way to new methods for political opinion formation and as such, a new angle for disinformation that had previously been observed in the analog world. First recorded allegations of such influence campaigns reach back to 2003. The claims were based on their observation that until 1998, contents on Russian forums had a predominantly liberal-democratic view (about 70%), but after 2000 a sharp surge of "antidemocratic" posts inverted the ratio to about 60 - 80% of "totalitarian values" [PKL03]. A temporal correlation exists with Vladimir Putin's rise to power in 1999 by appointment to the office of Prime Minister of Russia, promotion to Acting President of Russia after Boris Yeltsin's resignation and subsequent win of the Office in 2000. However, this correlation does not necessarily prove causation, as the Russian general population gained internet access during that timeframe, hence the possibility that the pre-2000 internet userbase simply not being representative of the general population or alternatively the population undergoing that transition in political views in that timeframe.

This phenomenon of pro-government web influence campaigns would later be dubbed as web brigades. Allegations and traces of such campaigns conducted or sponsored by Russian government entities have remained a topic since then. In 2012, a Russian hacker group published e-mails they claimed had been from correspondence by Nashi, a youth movement in support of Vladimir Putin and his ruling party United Russia [Eld12]. The material disclosed raised several allegations of paying for pro-Putin or anti-opposition contents. The leak contained price lists for bloggers and commenters who left comments on stories with negative content about Putin - up to 600,000 rubles for "hundreds of comments". Further paid content included Nashi activists to dislike anti-government videos on YouTube. Additionally, smear campaigns regarding prominent opposition figure Alexey Navalny, who was to be likened to Adolf Hitler [Eld12]. Nashi, however, denied these allegations.

Similar allegations have also been brought forward in a larger scope starting in 2014 during the Ukraine conflict. The conflict had started with the Euromaidan, a wave of increasingly large demonstrations after the implementation of the EU Association Agreement between the European Union and Ukraine was halted by Ukrainian President Viktor Yanukovych, who had been a supporter of the pro-Russian course and membership in its customs union. Yanukovych was eventually impeached, and Ukraine returned to a pro-European course. The conflict heightened tensions between pro-Russian and pro-government camps in Eastern Ukraine, where a large population of native Russian speakers lived. The conflict ended in the Russian annexation of Crimea and the civil war in Eastern Ukraine which was escalated by the Russian military intervention [SP16].
This conflict led to international sanctions, as the intervention in Ukraine and the annexation of Crimea were considered violations of international law. In order to influence public opinion about Russia's role in the conflict - in Russia, Ukraine and around the globe - these internet brigades were supposed to win public support of lifting sanctions and push the narrative of Russia not being the aggressor [Sin14]. In 2014, according to leaks these trolls had already been employed in comment sections of American media outlets. Some of those had featured bad English skills - a circumstance

that would later change with the professionalization of the IRA, one of the private contractors executing these campaigns for the Kremlin [Sed14]. The IRA would become a globally known entity due to its role in the United States election of 2016, as introduced in Chapter 2.2.2.

2.5 *Operation methods/techniques*. All of JTRIG's operations are conducted using cyber technology. Staff described a range of methods/techniques that have been used to-date for conducting effects operations. These included:

- Uploading YouTube videos containing "persuasive" communications (to discredit, promote distrust, dissuade, deter, delay or disrupt)
- Setting up Facebook groups, forums, blogs and Twitter accounts that encourage and monitor discussion on a topic (to discredit, promote distrust, dissuade, deter, delay or disrupt)
- Establishing online aliases/personalities who support the communications or messages in YouTube videos, Facebook groups, forums, blogs etc
- Establishing online aliases/personalities who support other aliases
- Sending spoof e-mails and text messages from a fake person or mimicking a real person (to discredit, promote distrust, dissuade, deceive, deter, delay or disrupt)
- Providing spoof online resources such as magazines and books that provide inaccurate information (to disrupt, delay, deceive, discredit, promote distrust, dissuade, deter or denigrate/degrade)
- Providing online access to uncensored material (to disrupt)
- Sending instant messages to specific individuals giving them instructions for accessing uncensored websites
- Setting up spoof trade sites (or sellers) that may take a customer's money and/or send customers degraded or spoof products (to deny, disrupt, degrade/denigrate, delay, deceive, discredit, dissuade or deter)
- Interrupting (i.e., filtering, deleting, creating or modifying) communications between real customers and traders (to deny, disrupt, delay, deceive, dissuade or deter)
- Taking over control of online websites (to deny, disrupt, discredit or delay)
- Denial of telephone and computer service (to deny, delay or disrupt)
- Hosting targets' online communications/websites for collecting SIGINT (to disrupt, delay, deter or deny)
- Contacting host websites asking them to remove material (to deny, disrupt, delay, dissuade or deter)

**Figure 2.2:** Excerpt from the report "Behavioural Science Support for JTRIG's (Joint Threat Research and Intelligence Group's) Effects and Online HUMINT Operations", leaked by Edward Snowden and published by The Intercept [Dha11].

Traces of influence campaigns conducted or sanctioned by Western democracies can also be found. In the leaks disclosed by whistleblower Edward Snowden it was shown that British intelligence agency GCHQ engaged in such activities as listed in Figure 2.2. These activities included influencing online discussions or discredit targets of agents as well as disseminating propaganda on social networks. The activities were not limited to targets abroad but also included domestic activities. The Joint Threat Research Intelligence Group (JTRIG), according to one leaked report, cooperated with security agencies as well as police units within the United Kingdom, with key objectives revolving around "denying, deterring or dissuading" criminals and 'hacktivists' as well as "deterring, disrupting or degrading online consumerism of stolen data or child porn' "[GF15]. Targets, however, also included the Anonymous hacker vigilante movement. While it can be stated that generally, those influence campaigns were geared towards crime prevention and not boosting the ruling party, the self-described playbook of methods and techniques in use by 2011 certainly involves a wider scope and larger audience than could be naively assumed. Figure 2.2 shows this list, which includes creating fake Facebook groups and Twitter accounts to influence discussions on topics, and doing

so with false personalities to promote those views as coming from real people - actions which mirror techniques that would be discovered as the keystone of Russian social media disinformation campaigns. Far more invasive and targeted techniques such as spoofed e-mails or delivering spoofed content on websites were also described (see Figure 2.2), essentially controlling the information targets can receive, as well as disrupting communication by denial of service of phone or internet connections. Also included in the report are needs for behavioral science insights - mostly social psychology - to improve their influence campaigns [GF15].

Likewise, the United States, for example, also has a campaign under the name of Operation Earnest Voice, whose aim is to use fake accounts to disseminate pro-US propaganda on social networks, with its target audience being abroad. The responsible United States Central Command (CENTCOM) claimed Facebook and Twitter were not targeted - with domestic propaganda being prohibited by provisions of the Smith–Mundt Act[1]. At the same time, the same bill allowed for such foreign content to be distributed domestically by state agencies. To conduct this operation, a 2.8 million USD contract [2] was signed that would allow government agents to post content to "foreign-language websites" was signed to provide for specialized software to create and maintain fake accounts to be used, integrate a randomized VPN connection and the ability for accounts to appear from anywhere [FC11].

## 2.2 The US Election of 2016

Possibly the most infamous target of wide-spread disinformation campaigns is the United States elections of 2016, most notably the Presidential election of that year, hence the focus for this thesis lying on this example. After a long campaign stretching two years, Donald Trump achieved an upset win over Hillary Clinton, taking back the White House for the Republican party after eight years. The election was particularly close, owing to the system of the electoral college. Instead of accumulating all votes and deciding the winner by selecting whoever receives the most across the country, each country has an allocated number of electors which are chosen by the voters in that state. Since most states have a winner-takes-it-all rule in place, the popular vote share and the corresponding share of electors can deviate significantly. Although Clinton received 2.5 million votes more than Trump (with 48.0% vs 45.9%), she still ended up losing the contest. By carrying the states of Pennsylvania, Michigan and Wisconsin with a total margin of less than 80,000 votes, he was able to achieve victory in the Electoral College.

This, however, is not even the closest nationwide contest in the history of the US. The 2000 election between Al Gore and George W. Bush was decided by a mere 537 votes in the state of Florida, the state that was able to tip the balance in the Electoral College either way - ultimately giving Bush a narrow edge. The ensuing conflict over the triggered recount was decided by the United States Supreme Court by halting all recounts and with Bush winning in the ruling of Bush v. Gore - though there is still a mixed view on who would have won if a recount had been completed, depending on the method and scope [Pay15].

[1] The Smith–Mundt Act was amended by the Smith–Mundt Modernization Act of 2012, which weakens some restrictions of domestic dissemination of such media intended for abroad audiences.

[2] The original solicitation notice from 2010 can still be found via Wayback Machine under https://bit.ly/2XTelEe (shortened link)

One major specialty about this electoral system stands out. Only a select amount of states - the so-called Swing States - are of real interest for political campaigns because the others are so solidly liberal or conservative that trying to switch the majority vote in those is usually a fruitless endeavor either way - which leads presidential candidates to disproportionally focus on those deciding votes [Str08]. Ultimately, this complicated electoral system can give electoral campaigns a way to circumvent the popular vote and still win a nationwide election. The Trump campaign strongly targeted the so-called Rust Belt, a region of concentrated steel industry in decline for decades by promising a steel industry comeback if he were elected. All three (swing) states mentioned earlier lie within that region. Therefore, it can be stated that very small shifts in votes - in the magnitude of thousands - at the right places can turn around the outcome of an election of over 100 Million votes in a tight election cycle [DPSV04]. Over the course of this chapter, an overview of the specific attempts of the Russian Federation to influence the outcome of that election will be given.

### 2.2.1 Timeline of Discovery

While most of the investigations and publications regarding Russian interference in the 2016 US elections were published after the election took place, first signs and investigations preceded election day by several months. On July 31, 2016, an official investigation codenamed Operation Hurricane was launched by the Federal Bureau of Investigation (FBI). While this investigation focused on possible ties between the Trump Campaign and Russia. Topics of focus were, for example, the hack of the Democratic National Committee (DNC), likely by Russian agents operating under the pseudonym “Guccifer 2.0” [Thi16], and the intentional and strategic release via Wikileaks.

This initial investigation kickstarted an ongoing process that is still not entirely concluded as of today. Although there are 17 known investigations into the matter, not all of those are of particular interest for the scope of this thesis, as some focus on possible ties between the campaign and Russian agents, the cyberattacks and possible obstruction of justice in an alleged coverup - all more or less related, but not at the core of the disinformation campaigns. However, the United States Senate Select Committee on Intelligence (SSCI) has released a partially redacted report that evaluates specific actions and methods the IRA used on Social Media [Uni19].

In October 2016, only weeks before the election took place, multiple agencies went public with claims that Russia had orchestrated the DNC emails hack and dissemination via Wikileaks. This happened in form of a joined statement of the U.S. Department of Homeland Security (DHS) and the Office of the Director of National Intelligence (DNI). In it, they claimed that the U.S. Intelligence Community (USIC) was confident that “the Russian Government directed the recent compromises of e-mails from US persons and institutions, including from US political organizations” [DHS16]. USIC is a group of intelligence-focused U.S. agencies from multiple federal organizations in several federal departments, whose head is the aforementioned DNI. Prominent member organizations are the Central Intelligence Agency (CIA), the FBI’s Intelligence Branch and the Office of National Security Intelligence. There are, however, many organizations under the Department of Defense involved as well, such as the Army’s, Air force’s and Navy’s respective organizations tasked with intelligence.

### 2.2.2 The IRA

The SSCI concluded that the IRA and its operatives overwhelmingly tried to influence US-American social media users. During its investigation, several relevant social media companies such as Facebook and Twitter cooperated with the SSCI to gain insight into those campaigns [Uni19]. The goal was polarization with usage of societal, ideological and racial angles. In its operations, the IRA mostly attempted to boost then-Candidate Trump's standing and to discredit Democratic candidate Hillary Clinton. However, social media operations also sought to boost support for third-party candidate Jill Stein. A possible explanation of that support is that by absorbing votes from the center-left spectrum, the chances of Hillary Clinton carrying the necessary states to obtain a majority in the Electoral College were reduced. In fact, Stein's vote share in key states such as Michigan was nearly five times as large as Trump's victory margin.

Previous actions furthermore involved interference during the Republican primaries to decrease chances of candidates with hardline stances towards Russia [Uni19, p.4-6]. Social media campaigns consist of multiple elements. Not only did seemingly private accounts of regular people interact with other users and share certain favorable or unfavorable political opinions, they also targeted the reputation of established media sources to erode trust in their validity. Advertisements were also used by Russian agents to spread political content, which was illegal under U.S. law.

#### Methods

The IRA is an entity based in St. Petersburg, Russia, specialized in influence operations domestically and abroad. Its operations have appeared since 2012, though the volume has increased substantially in recent years and become a high-profile player following its involvement in the 2016 U.S. elections [Uni19, p. 22-24]. Its domestic operations, e.g., spreading positive opinions about the state of economic recovery in 2015 and other pro-Kremlin content, are much less known internationally. It is being lead and funded by a Russian oligarch called Yevgeniy Prigozhin who has close ties to Russian president Vladimir Putin [Uni19, p.24].l In 2015, the IRA employed 400 people, who then were working 12-hour-shifts, under video surveillance [Che15]. These "trolls" were employed to create content on various social media platforms such as Facebook, Instagram and Twitter. Much of the known information stems from interviews with former employees, such as conducted by the New York Times in 2015.

According to ex-employee Ludmila Savchuk, employees used VPN services to assume different IP addresses as to not accidentally reveal the origin of the faux user accounts. These workers were given daily directives as to topics they were to engage about and had specific quotas to fill – specifying the amount of comments and political as well as other posts to be published on a shift basis. To amplify the effectiveness of their posts, they were to post supportive comments on the posts of other employees to make those more authentic (and more appealing to the platforms' algorithms) [Che15].

In order to gain followers and thus a larger audience, the IRA employed several methods [DI19] (at least for its German language accounts):

- Purchase of followers - Visible by sudden and erratic jumps in follower counts

- Fishing for followers - (Repeatedly) following and later unfollowing accounts in the hope that at least some follow back
- Narrative switching - Setting up an account to talk about all kinds of topics to get traction, then kicking off the agenda and move to the intended content (example of this can be seen in The Army of Jesus at Facebook, Figure 2.3

### 2.2.3 Facebook Operations

The IRA, as mentioned, has not only pushed for the creation of fake accounts to influence political opinion. The United States House Permanent Select Committee on Intelligence (HPSCI) published a set of political advertisements, commissioned by the IRA[3]. With over 3,500 advertisements and 470 identified IRA-created pages, more than 11.4 million U.S. users were exposed to their promoted content [Uni]. Organic content, meaning misinformation stemming from people reading the placed disinformation and sharing them, amounted to over 126 million U.S. based users being exposed to the content [Uni]. Several of the 470 placed fake pages stand out with the establishment of large gatherings of followers and specific target audiences, which we will analyze.

#### Example 1: African Americans

One such page was BM (Black Matters US). BM targeted African Americans with topics such as police brutality, racism and social injustice, thematically leveraging the Black Lives Matter movement in the United States and amassed over 200,000 followers by the time it was taken down. Another account, Blacktivist, was active in Facebook as well as Twitter and even managed to overtake the actual Black Lives Matter account in Likes, with 360,000 to 301,000 [OB17]. In its activities, it publicized multiple events during 2016, with some of those being organized by other, genuine groups in the first place but Blacktivist trying to increase turnout - and buying ads on the platform to increase their efforts. Their demographic targeting was sophisticated enough to, for example, specifically select the people in Baltimore, MD and Ferguson, MS - both of which are cities that had been the center of protests in conjunction with preceding police shootings [HGL+19, p. 21]. The goal here was not necessarily to steer voters towards candidate Trump but rather lower turnout by eroding trust in democracy. On one occasion, voting for Jill Stein was portrayed as a viable option and explicitly named as not being a lost vote - which had been a talking point of Democrats. [HGL+19][p. 9, 10]

#### Example 2: Evangelicals and Christian Right

Another well documented example for a group targeted by the IRA are Evangelical Christians, a loose movement within Protestant Christianity, and, more specifically, the Christian right [Cra19]. The Christian Right is broadly defined as the Social Conservative faction of Christians within the United States, with its core issues and stances said to be revolving around the following points [Cra19].

[3]The archives are available at `https://intelligence.house.gov/social-media-content/social-media-advertisements.htm`


**Figure 2.3:** Example: The Army of Jesus group on Facebook was created by IRA operatives, with the IRA themselves purchasing advertisements on Facebook to reach more users during the U.S. election cycle of 2016. Source: Facebook advertisements released by HPSCI

- LGBT+ topics (Opposition to LGBT+ rights progression)
- Education policy (creationism as official teaching, mandatory school prayer, homeschooling)
- Restriction of sex education and pornography
- Contraception and reproductive rights (ban of abortions)
- Reduction of Separation of church and state in general
- The perception of (Evangelical) Christianity as moral supremacy

The Religious Right is not a new phenomenon or group in American politics - it has been around in its roots for many decades, and its influence has been seen as growing since the 1970s, especially after the principal legalization of abortions by the Supreme Court ("Roe v. Wade", 1973). Its overall political views align broadly with those of the "Conservative wing", and they have become a force within the Republican party [Cra18].

An example for the targeting of this group by the IRA can be seen in Figure 2.3, in this example by the large Army of Jesus Facebook page, which tried to spin an angle that Hillary Clinton was tied to "Satan" and that her winning would mean that Satan wins. In fact, the Army of Jesus page was specifically noted in the SSCI report with its characteristic and representative "pattern of character development, followed by confidence building and audience cultivation, punctuated by deployment of payload content is discernable throughout the IRA' s content history" [Uni19, p.33]. While the targeting of African-Americans, as indicated in Chapter2.2.3, appeared to have its focus on disillusionment with Clinton and lower voter turnout for this Democratic voter group, targeting Evangelicals tended to attempt increasing turnout and support for Trump.

#### HPSCI Facebook data

In 2018, the HPSCI released over 3,000 advertisements Facebook identified as bought by the IRA during the 2016 U.S. election cycle[4]. Those archives - split into quarterly archives between the 2nd quarter of 2015 and the third quarter of 2017 - contained said advertisements in PDF format. While Twitter released tweets by the accounts they identified as belonging to the IRA in a machine-readable format, these PDFs do not qualify as such.

However, in [HGL+19], Howard et. al analyzed all pdfs released to them. They also signed a non-disclosure agreement with the SSCI during the timeframe they worked on their analysis. They claimed that their "analysis is the first, most comprehensive analysis of the data provided to the Senate by the social media firms". The HPSCI released the advertisements shared by Facebook with Congress, but did not enclose "the 80,000 pieces of organic content shared on Facebook by the IRA", whereas organic content denotes content shared on the platform that was not paid for.

In Figure 2.4, the amount of IRA Facebook ads by months can be seen from mid-2015 up to about fall of 2017. Several interesting things can be observed on this diagram: Rather modest activity started to constantly increase during the U.S. primaries in 2016, especially between the start of the Iowa caucuses (the traditional beginning) and a peak a month before the primaries ended. Activity started increasing again in time for the National Conventions, events both the Republican and Democratic Party hold where the delegates determined in the primaries formally nominate the candidate for the respective party in the upcoming Presidential elections. The sharp increase came to an abrupt stop by November 2016, when the elections occurred.

Spring 2017 saw another resurgence (see Figure 2.4), when the election was long settled, and Donald Trump already inaugurated. At the time, the Trump administration was dealing with an increased pressure of the Russia investigations that had started in the previous year. Michael Flynn had been named and resigned as National Security Advisor to Trump due to irregular contact with Russian

[4]The archives are available at `https://intelligence.house.gov/social-media-content/social-media-advertisements.htm`

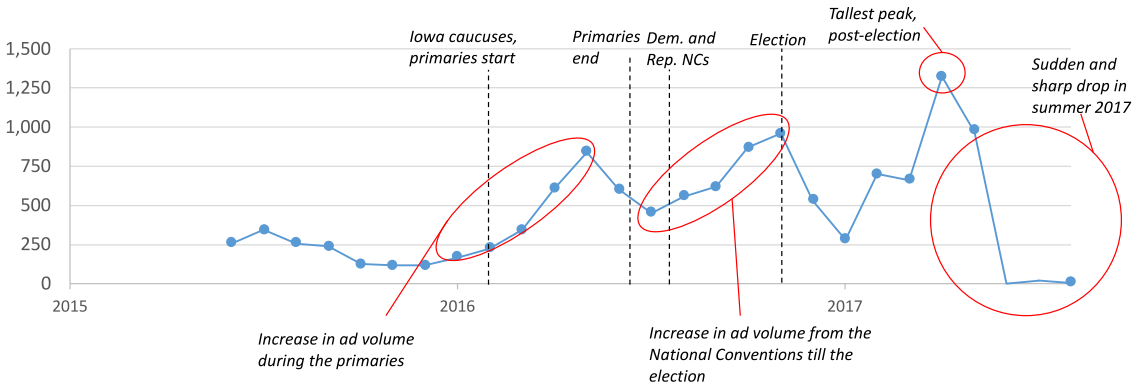


**Figure 2.4:** Amount of IRA advertisements on Facebook, by month and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.15]

officials, the evolving HPSCI investigation into the Russian interference and finally the dismissal of FBI director James Comey increased public pressure. These developments, with emphasis on the last, led to the appointment of Robert Mueller as Special Counsel to independently investigate all Russian activities (and the administration's handling of those) in the 2016 U.S. elections [RL17].

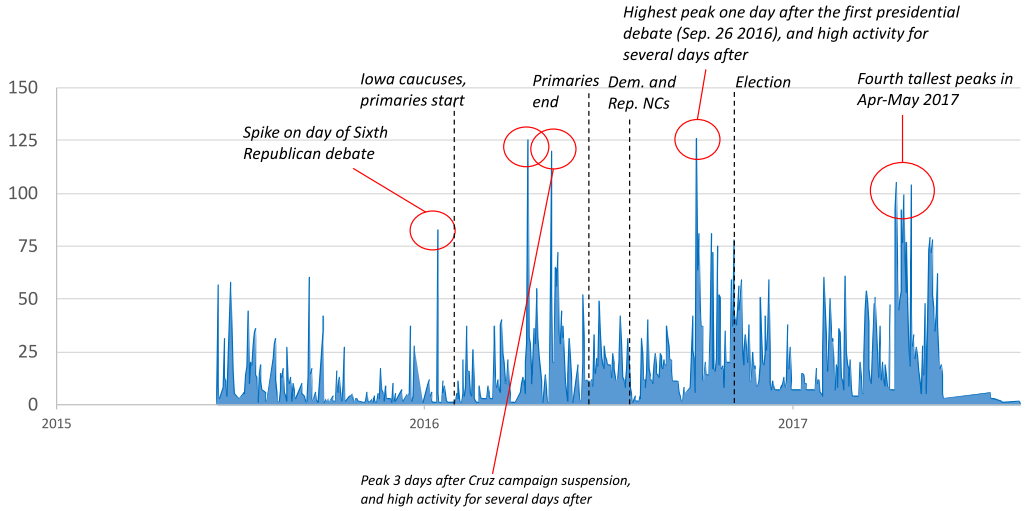


**Figure 2.5:** Amount of IRA advertisements on Facebook, by day and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.15]

The same data is analyzed in a more fine-grained, daily total graph in Figure 2.5. The general observations are similar as in Figure 2.4: Activity peaks appeared in time for the sixth Republican debate and again during major campaign events, such as Ted Cruz's campaign suspension. The highest recorded daily peak coincided with the first presidential debate in September 2016. As daily activity ramped down after the election, isolated high advertisement activity in Spring of 2017 occurred afterwards. As hypothesized in the previous paragraph, this might be related to the pressure on the Trump administration by congressional investigations during that timeframe.

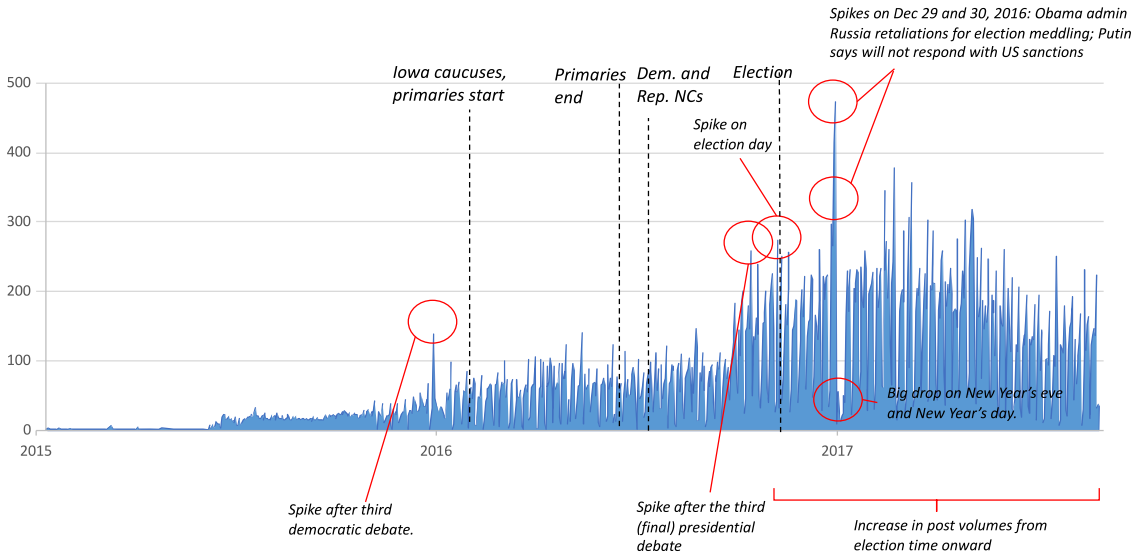


**Figure 2.6:** Amount of IRA posts on Facebook, by day and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.16]

Apart from the paid content pushed by the IRA, there was, of course, a much bigger group of content: All sorts of posts created by IRA-controlled accounts. The distribution of activity looks different in this category, as displayed in Figure 2.6. The second half of 2015 showed a relatively constant amount of activity, with the first distinct spike just after the third Democratic debate (December 19th, 2015). From there on out, daily totals became increasingly erratic, but spikes became significantly higher as they had been before. Overall, the first half of 2016 showed a modest volume of activity, followed by the amplitude of spikes rising by a factor of two. The final presidential debate and election day showed strong spikes, but the largest singular spike occurred on December 29th, 2016. On that day, the outgoing Obama administration announced a set of retaliations for what they alleged to be election meddling by Russia, most prominently the expulsure of 35 Russian diplomats [San16]. Notably, the volume of organic posts stayed on persistent high volume throughout the spring of 2017, generally exceeding the time leading up to the actual election itself. The volume then again decreased afterwards, during the course of 2017[5].

### 2.2.4 Instagram Operations

Previously, mentions of social media activities by the IRA have been contained to Twitter and Facebook, the social network. Facebook Inc., however, is also the owner of Instagram, which it acquired in 2012. While content on Facebook is more focused on the users publishing posts of different types of content - such as photos or videos with captions, text posts, polls - Instagram puts photos and photosets at the center of the content stream. Users are, however, able to make the content searchable by appending hashtags or the location, through which they can be found via the search function. Figure 2.7 displays the daily total posts on Instagram by accounts attributed to the IRA, again made available via the HPSCI. One discrepancy to Facebook's daily post totals, as seen

[5] Again likely related to the IRA activities being uncovered by Facebook as the year went on

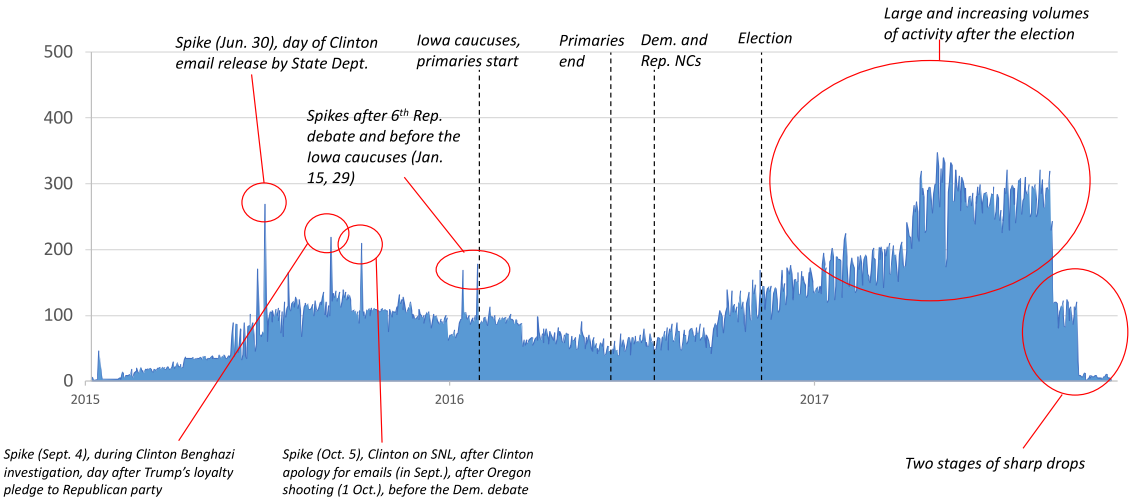


**Figure 2.7:** Amount of IRA posts on Instagram, by day and key 2016 U.S. election cycles events marked. Image source: [HGL+19, p.16]

in Figure 2.6, is quite interesting. While Facebook had a tendency of rising spikes, with a constant low level in 2015, increases in 2016 and a maximum in early 2017, Instagram shows a different picture: after a short phase of initial increases from a very low level, several high spikes occurred in 2015 very early in the wider 2016 election cycle. Most notably, the biggest total spike until well into 2017, was at the time of the June 30th, 2015 partial release of Hillary Clinton's emails (see Figure 2.7). The emails became of interest after it had become public that she had used a private, not properly secured server for her official communication (which in turn became a common attack vector for Republicans during the 2016 election campaign). At the time, the e-mails were released in monthly batches after a court ruling.

As with the Facebook IRA activities, spikes were also observed at the Republican presidential debates and at the beginning of the Primary election cycle with the Iowa primaries. Another interesting spike observation correlates with the September 4th, 2015, Benghazi investigation. The investigation revolved around the attack on the U.S. embassy in Benghazi, Libya, in 2012. The U.S. embassy was attacked by Islamic militants, which caused the death of 4 U.S. and 7 Libyan citizens. During that time, Hillary Clinton was the Secretary of State of the United States, and therefore responsible for the U.S. embassies and diplomatic staff.

The Benghazi attack was a prominent attack vectors against President Obama and Secretary Clinton (as well as National Security Advisor Susan Rice), with common Republican talking points revolving around negligence and wrongdoing. In total, 10 investigations were conducted, but no wrongdoing by the mentioned three high-level officials was found, though lower-level officials had been criticized by not responding to requests for additional security measures. Regardless, Benghazi remained a Conservative rallying-cry against Hillary Clinton for the 2016 election campaign [OTo16].

Early 2016, however, saw a decline in activity on Instagram, though the trend stopped and reversed after the Democratic and Republican National Conventions. Activity steadily rose until the 2016 election, and far beyond that. As with Facebook, higher levels than in the run-up to the election were measured in Spring 2017, reaching its maximum around an April-May 2017 timeframe. Afterwards,

activity sharply dropped in two phases to near-zero. This again indicates that the IRA influence was never meant to be contained to the 2016 election but was meant to be applied on a longer timeframe.

### 2.2.5 Twitter Operations

During hearings in the U.S. Congress in the fall of 2017, Twitter had committed to providing updates on its investigation on election interference conducted on its platform [Twi18]. During its review, Twitter identified 3,841 individual accounts tied to the IRA which accounted for a total of about 175,000 tweets, though only 8.4% were election related [Mue19, p. 28-29]. However, a much more sizable amount of automated accounts (bots) were discovered in the process as well: In its January 2018 report, Twitter named a total of over 50,000 accounts tied to Russia's disinformation campaign[Twi18].

Twitter has expanded its search for state propaganda on its platform to various accounts and campaigns around the globe – including the European Union. All results, including data archives, are preserved and publicized on Twitter's Elections Integrity Hub. These exposed accounts and its removed but archived accounts and tweets originated, for example, from Venezuela (1,196 apparent state-backed influence campaign participants, 764 with likely but non-definite ties), Iran (4,779 accounts associated or backed by the Iranian government), China (936 accounts tied to deliberately spreading discord tied to Hong Kong) or even Spain (130 accounts tied to the Catalan independence movement)[6].

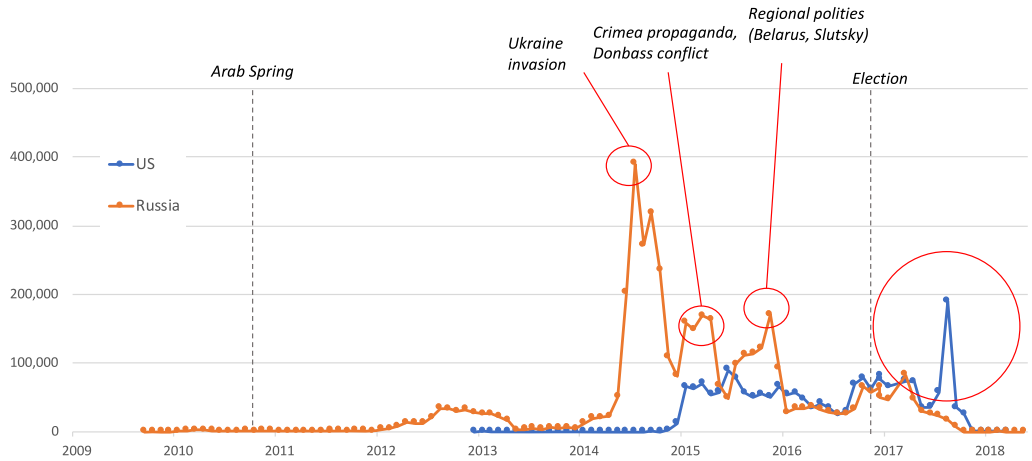


**Figure 2.8:** Amount of IRA tweets on Twitter targeted at Russia (orange) and the United States (blue) between 2009 and 2018. Image source: [HGL+19, p.28]

As Figure 2.8 displays, Russian activities were the prime focus of the IRA up until 2016, when U.S. activities started to generally receive a higher volume of tweets. Russian activities followed a similar schema as in the U.S. with big and significant events drawing spikes in activity. Most notably, and strongly outperforming any U.S. activities, the start of the Russian military intervention

[6] All Twitter data archives are available at `https://transparency.twitter.com/en/information-operations.html`

in Ukraine in 2014 caused the largest spike in activity. Other significant events pertaining to that, such as the Crimea occupation and the heating conflict in Donbass, Ukraine, led to further spikes. The time around the U.S. election and mid-2015 showed highest average activity in the U.S., though as on Facebook, Twitter activity saw the unusual Spring 2017 spike as well, just before activity dropped sharply [HGL+19, p.28].

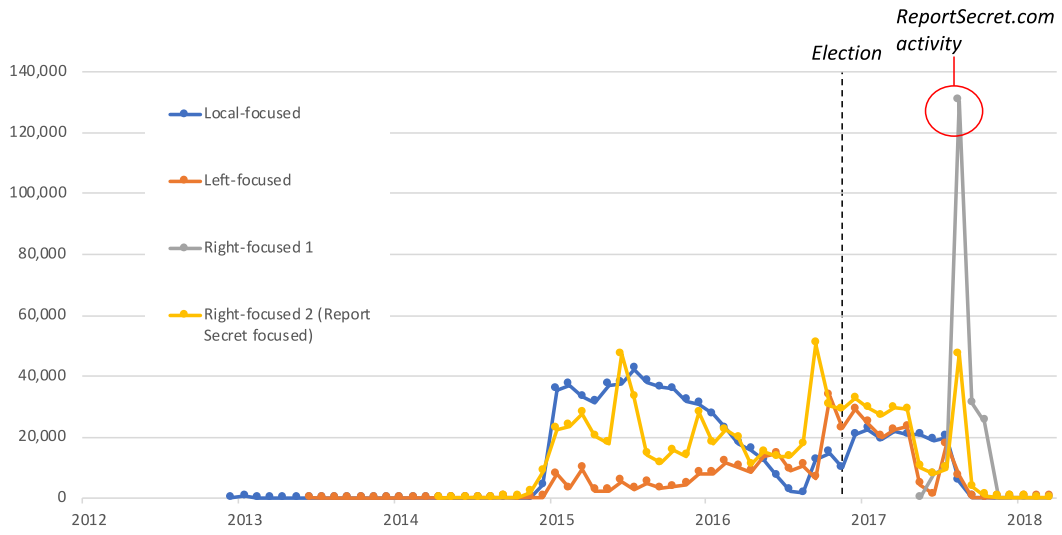


**Figure 2.9:** IRA Tweets on Twitter with different types of focus - regarding local events or political ideologies between 2013 and 2018. Image source: [HGL+19, p.28]

Figure 2.9 displays IRA activity divided in categories, split by political ideologies (left / right) as well the categorization of local focus. Tweets on the right are hereby almost always outranking tweets on the left, and especially in 2015 and early 2016, the IRA had a local focus rather than targeting right-wing or left-wing audiences. It is, however, again easy to see that the IRA focused on right-wing audiences in the U.S. while also targeting the opposite [HGL+19, p.28].

An interesting finding raises the question about a potential flow of knowledge from Wikileaks to the IRA. On October 7th, 2016, Wikileaks released a set of emails obtained in a hack earlier that year[7]. Podesta was the campaign chairman of Hillary Clinton's 2016 presidential campaign. This gave the public a look into the inner workings of the Clinton campaign and its correspondence with figures in the Democratic National Committee (DNC). Figure 2.10 shows the activity of left-wing and right-wing troll accounts during that time frame. Very prominently, about 35 to 30 hours before the e-mail drop, accounts in both ideology audiences started an activity burst. That burst slowly declined and was abruptly finished with a strong peak about 10 hours before. It is not known whether this was a coincidence or not.

[7] Which is not to be confused with the hack of the Democratic National Committee, whose e-mails were given to Wikileaks and later released also.

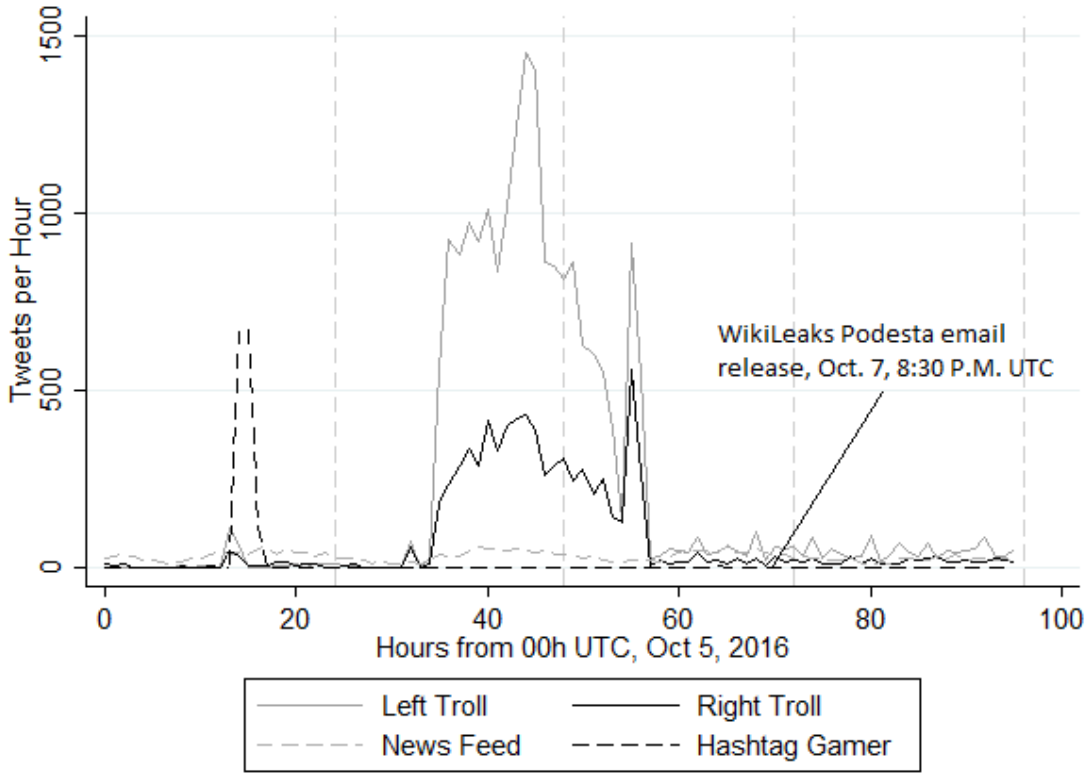


**Figure 2.10:** IRA Number of tweets per hour starting from October 5th, 2016, 12:00 AM UTC. Number of tweets by accounts categorized as politically left (grey) / right (black) trolls is plotted, with high activity appearing before the release of John Podesta's emails by Wikileaks. Image source: [BLW18, p.19]

As Figure 2.11 shows, overall, Twitter has been the prime medium of the IRA. On a yearly basis, monthly post averages did not show significant variation from 2015 to 2017, though 2017 might be affected by the steep decline during the year likely due to IRA accounts being discovered and suspended. On the other hand, Facebook showed a large increase from 360 monthly posts in 2015 up to about 4220 in 2017. Instagram, too, showed a large increase from about 2100 posts in 2015 to roughly 6000 posts in 2017. Despite those big increases, though, Twitter activity remained by far the highest target of the IRA.

### 2.2.6 Activity Drop of 2017 and After

As can be observed in analyses of all three Facebook, Instagram and Twitter, activity by the IRA increased after the election and peaked around Spring of 2017, only to be followed by sharp drops. It is possible that the rise in activity made it more obvious and detectible to the social networks, and that ensuing crackdowns were able to catch most of these accounts by the summer. It is, however, also entirely possible that the IRA moved its activities to a different set of accounts or applied different methods to avoid detection. Another possibility is that at least some of the resources have been shifted to operations in other countries than the United States.

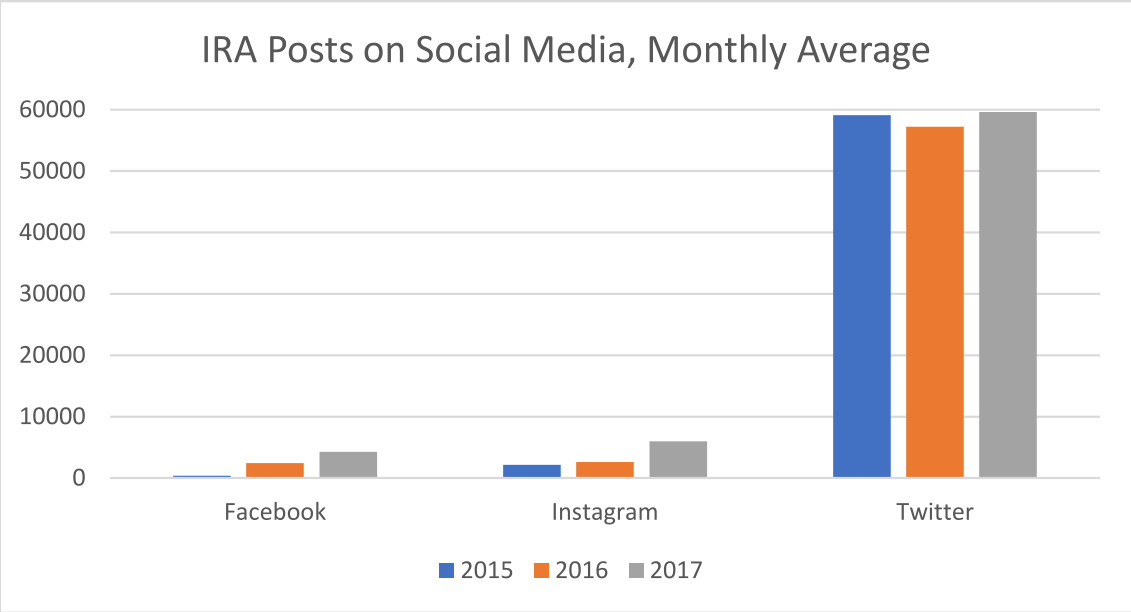


**Figure 2.11:** Monthly average amount of IRA posts on Twitter, Facebook and Instagram by year, from 2015 to 2017. Data source: [HGL+19, p.5]

In 2018, reports based on criminal charges filed by the U.S. against the IRA indicated that its budget allocated to its troll program had been doubled, with spending from January to June 2018 alone coming close to matching the budgets for the entire years of 2016 and 2017 [Hal18]. In November 2018, on election day for the U.S. midterm elections, it was also reported that the U.S. Cyber Command had launched a cyberattack against the IRA, to "prevent the Russians from mounting a disinformation campaign that cast doubt on the results" [Hal19].

In February 2020, warnings again surfaced about Russian ongoing interference in the U.S. elections, this time the November 2020 cycle. Intelligence officials had briefed lawmakers from the U.S. House of Representatives about a plot to reelect Donald Trump as President of the United States as well as interfere in the Democratic 2020 primaries [GBHF20]. In the same article, The New York Times also mentioned that according to officials, "search for issues that stir controversy in the United States and use various methods to stoke division", similarly to the 2016 interference. Whether or not such activities will be observed in the months leading up to November 2020 remains to be seen.

## 2.3 Efficacy of IRA activity

Until now, the IRA activity has been looked upon in aspects of how and what: How were campaigns conducted and what did they do. To properly evaluate the significance of their operations, it is crucial to assess the actual impact of their social media operations.

Surprisingly, there is not a lot of research available. That is astounding, considering the high resonance the IRA's activities have gotten in public coverage. Though at the same time, trying to estimate the psychological impact of the IRA's campaign is difficult as the easiest way - conducting

an own disinformation campaign specifically for research - is inherently unethical, if not illegal, and therefore cannot be executed. This point was also made by a study that recognized this and therefore moved on to different means: They used the (unrestricted) IRA Twitter data set to compile a list of targets and when they interacted with IRA-controlled accounts on Twitter [BGM+20]. They compared this with a 2017 longitudinal survey of Republican and Democratic users who reported to be active on Twitter. The information combined provides with subjects of either political party and their political attitudes before and after interaction with the IRA [BGM+20]. In the context of this study, an interaction with an IRA-controlled account may consist of liking or retweeting one of their tweets (or tweet they are mentioned in), as well as following their account[BGM+20].

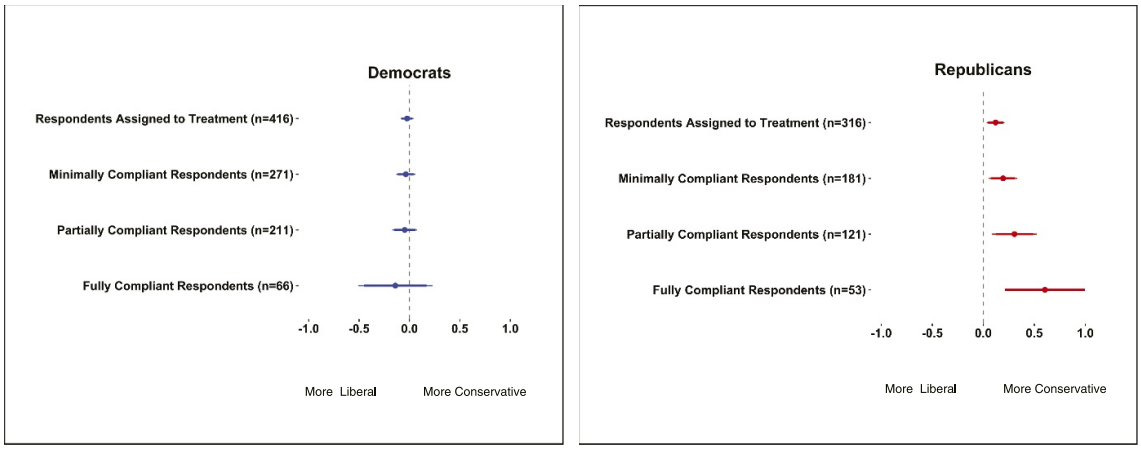


**Figure 2.12:** The measured effect of following the opposing-views bot on Twitter by the respondents. The dots represent the estimated averages, with the red (Republican) and blue (Democrats) lines denoting the 95% (bold) and 90% (slim) confidence intervals. Image source: [BAB+18, p. 9220]

The survey published in 2018 dealt with a similar, more generic, question: Will exposing Twitter users with Democratic or Republican views to opposing political content via a Twitter bot change their views[BAB+18]? To do so, they sought out Twitter users with a survey. For a financial compensation, they would provide information such as their Twitter ID, political views and (social) media habits before, a week later, they were assigned to Democratic and Republican control and treatment groups. The people in the treatment groups would then be following a bot posting liberal messages (retweets of officials or otherwise active influencers) for Republicans and conservative messages for Democrats, with those bots publishing 24 messages a day. One month later, all respondents - control and treatment - would then fill out another survey to measure the changes in all four groups [BAB+18].

What they found was that Democrats had a slight - though not statistically significant - shift towards a more liberal political view, as Figure 2.12 shows (left). Vice versa, Republican respondents (right) showed an increased conservative view after the study concluded - and for them, the effect was much more distinct. Ultimately, both partially and fully compliant Republican respondents showed a significant increase of Republican views. The results also showed that the higher the level of compliance by the respondents, the stronger the shift was.

The study [BAB+18, p.9220] therefore showed that being exposed to opposing view political content by bots can lead to paradoxical increase of one's own belief, increasing political polarization. This is an interesting result, as a reasonable assumption is being exposed to certain content will lead to accept it as a truth. This result indicates that the effects of specific social media content exposure are not that obvious. However, those results are only partially applicable to the concrete question of how IRA activities might have influenced voters.

1. IRA-exposed subjects were not aware of the fraudulent accounts they encountered
2. Subjects might have encountered multiple sources with multiple political objectives rather than just liberal / conservative content spread

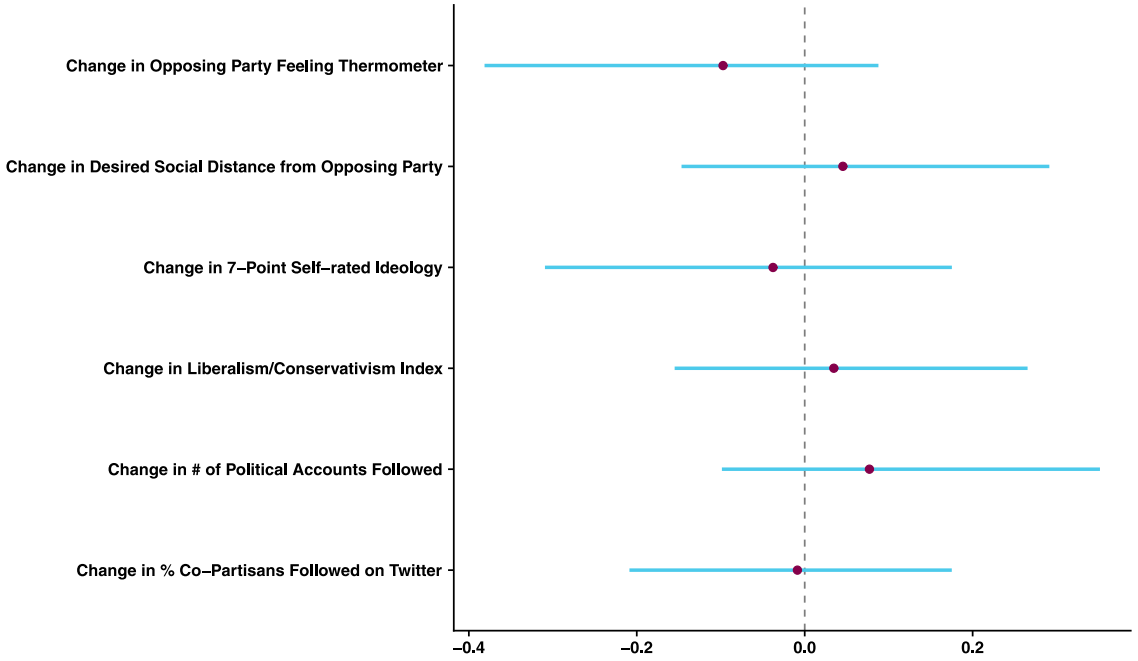


**Figure 2.13:** Results of interaction with IRA-related tweets over time - the points correspond to average effect of exposure, and the blue lines the 95% credible interval. Image source: [BGM+20, p. 246]

To deal with the first issue, they used the survey data and the IRA Twitter archive to look for effects of this hidden IRA-controlled content exposure. By utilizing this pre-existing data, several shortcomings were inherently present. First, there was a specific, narrow window in which the study data provided data on change in attitude. Second, towards the end of the U.S. election cycle, Twitter started becoming increasingly aware of IRA activity on its platform as the media started reporting on U.S. authorities' suspicions and ongoing investigations. Therefore, many IRA accounts started getting suspended throughout early 2017, reducing the sample size during the course of the study in 2017. As per their measurement, 75% of IRA interactions by their survey's respondents had happened with accounts that were already suspended by the time of the suspension wave [BGM+20, p. 246].
A possible effect on the results might be a kind of selection bias: As the IRA accounts with the biggest follower counts - therefore more successful - were suspended relatively early, they could not have interactions with the study participants. Thus, the sample of IRA accounts still active in the

October - November 2017 timeframe cannot be viewed at as inherently representative of the IRA's total efforts, and specifically not of their actions during the 2016 election cycle, between about mid-2015 and November 2016. Third, the distinction in the study is only done in a binary fashion - group receiving treatment (1) and control group (0) rather than comparing treatment at different intensity and control group, which might lead to a different picture. Added to this, the treatment was non-distinctive regarding the type of interaction. It is possible that some direct interactions, such as directly tweeting at an account might have a bigger impact compared to less direct interactions as liking a tweet or getting followed by them.

However, with those three limitations noted, there were no findings that contact with IRA-controlled accounts influenced study participants, as per [BGM+20, p. 246]. Figure 2.13 visualizes 6 properties such as self-rated ideology, the liberalism / conservatism index or the change of views about the opposing political party. All six of those had minimal effects well below the significance threshold.

Bail et al. also obtained another dataset by pollster YouGov that measured self-reported ideology on a five-point scale from February 2016 to April 2018, which provided them with responses over a longer time-frame, as they wanted to investigate the effects of a longer-term influence vs. their month long study[BGM+20, p. 246]. For this dataset, they were able to distinct between all engagements of the users compared to only direct engagements (e.g. replies). There was no significant change in self-reported ideology scale for either of those types, for either 1+, 2+ or 3+ interactions with IRA accounts [BGM+20, p. 250].

In conclusion of the studies explored above, there is no evidence present that the IRA (dis-)information campaigns on Twitter in the United States 2016 elections had a significant impact on the election outcome.

## 2.4 Impact in the European Union

While the 2016 U.S. election has arguably been a high-profile case of foreign disinformation campaigns, allegations of comparable interference have been made in other countries as well. Such allegations have also been made regarding the United Kingdom's referendum on its membership in the European Union in June 2016, which predates the U.S. election (see Chapter 2.4.2). Only months after the November 2016 U.S. election, while the U.S. was already publicly debating it having been a target of social media campaigns, the French presidential elections of 2017 experienced similar patterns of social media activity, as we will now explore.

### 2.4.1 2017 French Presidential Election

In 2017, France entered a polarized election season. Incumbent president François Hollande did not run for reelection. The election came down to a contest between the eventual winner Emmanuel Macron and right-wing candidate Marine Le Pen. Two days before the final vote, the so-called "#MacronLeaks" were released and quickly gained large coverage by the media and started trending on Twitter, after being shared by WikiLeaks [Moh17]. In its timing, the leaks were released just ahead of the media blackout that French electoral law imposes for 44 hours before an election.

The leaks consisted of 21,075 emails by the Macron campaign, which had been hacked between late April and early May, and uploaded to Pastebin, an online platform where users can anonymously upload snippets of text. The data was then posted to 4chan, an online message board, as well as Twitter, and mostly (rapidly) shared by alt-right[8] figures. Wikileaks gave the release traction as well by mentioning the leak multiple times [Mar17].

Some accounts on social media claimed that the leaks contained proof of wrongdoing such as tax evasion - claims that were not substantiated afterwards [Gre17]. No criminal wrongdoing seems to have been discovered and published. For this operation, however, this would not have been necessary. With such a specific timing right before the media blackout, the rumors of such wrongdoings being proven by the drop could have spread widely without the ability for new developments to be spread by the media.

After the election, researchers started analyzing the social media activities surrounding MacronLeaks. Indeed, bot activities indicating a disinformation campaign could be found: one study by Emilio Ferrara aimed to analyze Twitter activities and uncover bots participating in tweeting about it, as well as qualifying intentions of a possible disinformation campaign [Fer17]. They found 99,000 users identified as participating in MacronLeaks discourse. Of that, 18,324 were classified as bots (about 18%) [Fer17, p. 8]. An interesting observation also revolves around the discrepancy in temporal activity between generic election related tweets and MacronLeaks discussion, as illustrated in Figure 2.14.
While the grey line, the generic tweets, follow a pattern of day and night cycles (with late-night hours showing very low activity) and peaks at typical events sparking more discussions (e.g. debates), comparable to a sinus function. Meanwhile, the curve is entirely different for MacronLeaks-associated content. The sudden onset and distinctive peaks are typical for disinformation campaigns, especially due to not aligning with the circadian pattern of the French electorate [Fer17, p. 7]. On May 3, 2017, a user on 4chan hinted at a set of hacked documents regarding Macron, also contributing to speculation and activity before the release even occurred [Moh17].

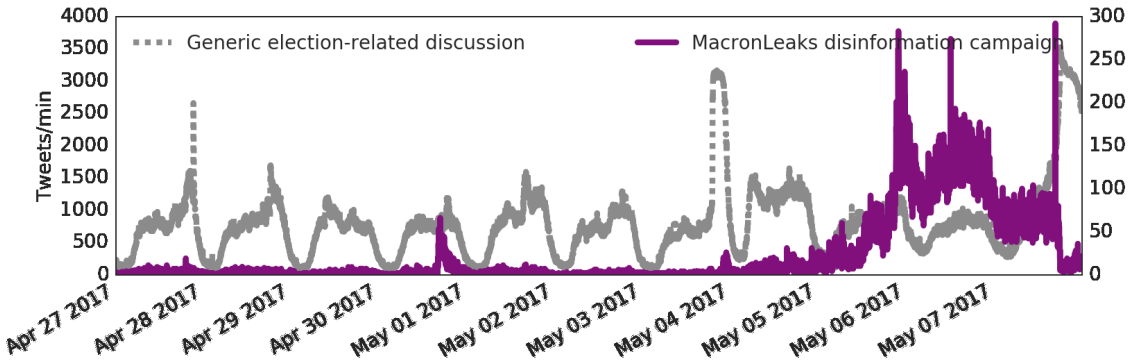


**Figure 2.14:** Generic election (grey) and MacronLeaks (purple) related tweets during April 25th and May 7th, 2017. The tweets counted were related to the French election campaign and MacronLeaks. Time in UTC. Image source: [Fer17, p.29]

[8] Self-identification of a group of anti-mainstream U.S. conservatives, associated with ideas of white nationalism, as per Merriam-Webster

A curious finding revolved around past activities of the bot accounts identified: Some accounts had supported Donald Trump in the year prior and went silent after the election just be woken up before the French election and then tweeting about MacronLeaks[Fer17, p.3]. This gives support to the hypothesis to some coordination of whoever controlled those accounts, possibly as a sort of black-market ware, or the same power trying to support both Donald Trump and Marine Le Pen. The general connection between the "alt-right" and MacronLeaks also became underlined by "MAGA" and Trump being the top-two terms in the profile descriptions of accounts involved in the election and MacronLeaks discourse [Fer17, p. 26], thus implying a big part of the MacronLeaks audience actually being Trump supporters instead of the French electorate. At the same time this could explain the failure of this campaign to make an impact: As a sizable amount of the audience appeared to be Trump supporters and thus American (at least the human accounts), their opinion was without consequence at the polls, as no vote could be cast by most of them.

### 2.4.2 Brexit

Skepticism about the UK's membership has been present in the British population since before it joined the European Communities (EC) in 1973 and has remained until today [US13]. In 1975, a referendum after renegotiated rules of the UK's accession to the EC fell clearly in favor of remaining. However, opposition to EC and later European Union (EU) membership never dissipated. The UK Independence Party (UKIP) under Nigel Farage kept mounting pressure, as did hardliners in the Conservative Party. In the manifesto for the 2010 election, in which the Conservative party would displace the Labour party from power in a coalition with the Liberal Democrats, they were committing to an "in/out" type of referendum. UKIP's strong election performances arguably pressured Prime Minister of the United Kingdom, David Cameron to pursue a referendum, which he did by getting legislation passed for a referendum by the end of 2017 [Ush16]. As the Conservatives gained an absolute majority in the 2015 election, the referendum ultimately took place on June 23rd, 2016.

Prior to the referendum, Cameron entered negotiations with the European Union to get changed terms for its continued membership in the European Union additionally to its opt-out clauses it had gained during entry in the 1970s. Cameron's demands revolved mostly around opting out of the "ever closer union" - the EU's vision of continued member states integration with each other - as well as limitations to free movement and EU nationals receiving child benefits from the UK.

David Cameron had made clear that he would seek certain reforms in the UK's membership duties in the EU and, given he were able to reach those, he would support remaining in the Union. Eventually, the UK and EU-27 did agree on a deal revolving around the issues mentioned above in February 2016, and Cameron went on to campaign for Remain [Wri16]. Despite this, the referendum fell in favor of the Leave campaign.

Around the time when allegations of Russian disinformation campaigns rose in the United States, the same applied for Brexit: allegations of bots used to influence public opinion [HK16] []. On Twitter, multiple bots were involved. Interestingly, some of them appeared to be repurposed bots that had previously amplified other political messages, such as the Twitter account Col_Connaughton, that had been sharing pro-Palestine content and switched to support the Leave side in the Brexit referendum. It was classified as a bot intended to retweet other accounts and to thus increase their reach [HK16, p.2].

Overall, they estimated about 14% of all content under the most popular Brexit-related hashtags stemmed from bots [HK16, p.4], though tying those accounts to potential foreign campaigns (or disproving it) has not reached a conclusion. Some evidence does, however, exist. Even though there is no dedicated IRA Twitter archive for Brexit, some of the accounts that we discovered to be involved in the U.S. disinformation campaign also published Brexit-related content, implying at least some degree of interest by Russia [LCFH18].

The U.K.'s Intelligence and Security Committee of Parliament (ISC) investigated allegations of Russian interference in the Brexit referendum. Besides allegations of using Russian state television Russia Today (RT) to spread disinformation, claims of social media campaigns arose in the UK as well. It delivered its report to the U.K. government led by Prime Minister Boris Johnson in October 2019. The government, however, rejected calls for the report to be made public immediately and specifically before the parliamentary election later that year [Per19]. As of June 2020, this report remains unpublished.

As of now, the evidence of social media campaigns conducted by Russia appears to be significantly weaker than its counterpart in the US. Amplifying and providing one-sided coverage e.g. through state-controlled RT[9] is not necessarily disinformation but suggests at least some form of political interest in the referendum. While, however, the disinformation aspect is not conclusive - the picture might become clearer once the blocked report is released - another adjacent aspect is present in both the U.S. election and UK referendum: The application of big-data for voter targeting.

## 2.5 Big Data

Big Data describes the field of collecting, analyzing and processing sets of data which are larger than what can typically be processed by traditional approaches [DGG15]. Modern data collections have become larger and larger. An example for a large set of data would be the data sets of the US elections of 2012 and 2016. A great example for analysis of such scope is visible in Figure 2.15. It displays a map of the United States, its 50 members states and its electoral districts. For each county, the map contains an error, whose length is determined by the magnitude of voter shift between 2012 and 2016. Facing to the right (and being in red), it denotes gains for the Republican Party, while facing to the left (and being in blue), it does the same but for the Democratic Party. This visualization is perfectly suitable to underline a point made earlier, in Section 2.2. The numerous big red arrows in the mid-western US are centered in the so-called Rust Belt, a manufacturing-driven region prominent in - among other states - parts of Pennsylvania, Michigan and Ohio. Those three states were narrowly flipped by Donald Trump and significant in securing him the majority in the Electoral College. Hillary Clinton made gains centered in urban areas. However, not all applications are as educational or insightful as this example.

[9] Russia Today (RT) is a Russian, state-backed internal television network.

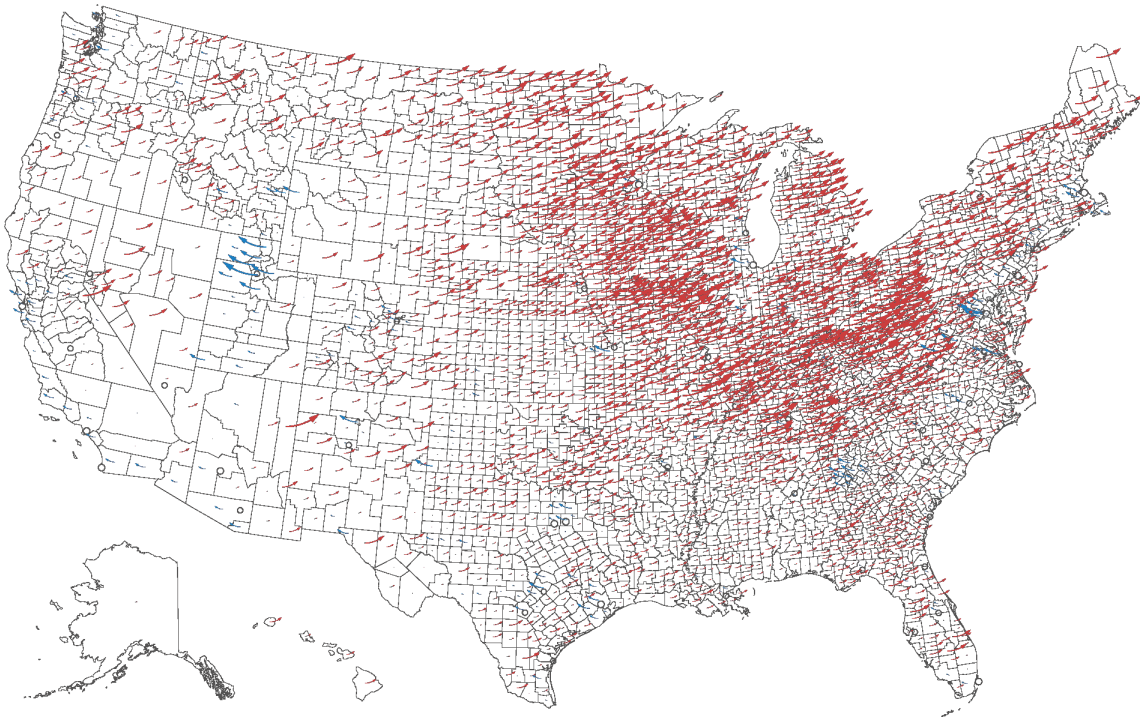

**Figure 2.15:** Map displaying the voter shift in each US district in the 2016 U.S. election. Source: Larry Buchanan et al. for The New York Times, based on election results data compiled by the Associated Press, color-adjusted by author for better visibility

### 2.5.1 Cambridge Analytica

A formidable example for the dangerous aspects of big data in election campaigning is the CA scandal of 2018. CA was a company founded in 2015, in London [VP15]. Its parent company, SCL Group, described itself as a "global election management agency". It came into media spotlight early on, as far back as July 2015, when Politico reported about its partnership to then-candidate for the US presidency, Ted Cruz [VP15].

Its portfolio has been described as providing "governments, political groups and companies around the world with services ranging from military disinformation campaigns to social media branding and voter targeting" [VP15]. Besides Robert Mercer, researcher into Artificial Intelligence, hedge fund manager as well as Republican donor in the 2016 election cycle, it was co-founded by Steve Bannon, widely known as former executive chairman of Breitbart News, a US-based website categorized as politically right-wing to far-right and later White House Chief Strategist for Donald Trump [Ill17].

CA's methodology itself is inspired by the works of Michal Kosinski et. al., who demonstrated the ability to predict political affiliation, sexual orientation or race based on Facebook Likes, psychometric tests and demographic information [KSG13]. To acquire the required information, CA utilized an app called "This Is Your Digital Life". The app provided a personality quiz for users, who would approve the use of their data for academic purposes. However, a design flaw in Facebook's API allowed them to also access the data of the participants' Facebook friends. As later admitted by Facebook, utilizing this method, CA collected data on up to 87 million users, about

70 million of those located in the United States [Sch18b]. Of those affected, collected information would include their profile, current city, birthday information as well as (page) likes. Even more severely, some users permitted This Is Your Digital Life to access their news feed, private messages or timeline and would end up getting analyzed even further [Lap18]. This would also affect their friends, as their private content would be available in the target's newsfeed and messages as well. With that information, enough data could be collected of those users and their friends to create psychographic profiles. This, in turn, was sufficient to make predictions as to what advertisements to target them with most effectively.

First allegations of illegal activities by CA were published in December 2015 by Harry Davies in The Guardian, according to uncited documents. Davies described the company creating psychographic profiles of US citizens, with the data being collected without permission through an online survey [Dav15]. The actual scandal itself emerged with leaks by the whistleblower Christopher Wylie, who had been an employee at CA in its earlier years from 2013 to 2014, about his time spent there. Wylie had already supplied a 2017 The Observer article "The Great British Brexit Robbery" with CA inside information as anonymous source, though its accuracy was at least in parts disputed. On March 17th, 2018, The Observer and The New York Times (NYT) jointly broke the main story of the CA scandal, referring to the tens of millions of user profiles being collected.

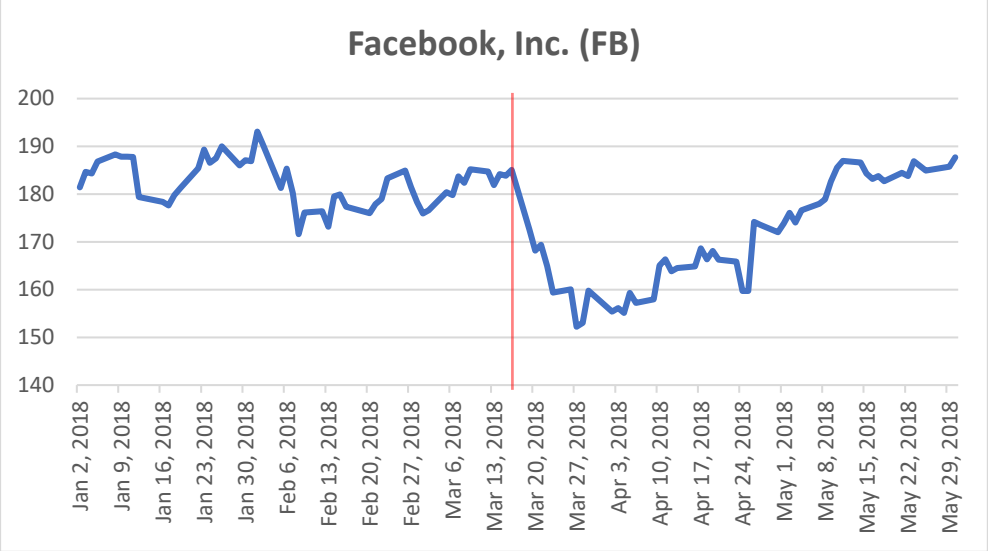


**Figure 2.16:** Development of Facebook's stock price between January and May 2018. The red line denotes day the CA story broke on the NYT and The Guardian. Data via Yahoo! Finance

For Facebook, the CA scandal was the kickoff to a difficult 2018 in terms of results and public scrutiny. Figure 2.16 illustrates the severe stock volatility introduced by the CA scandal. The major revelation occurred on March 18, with the last closing value denoted by the vertical red line. As analyzed by CBS News, Facebook's stock price fell by over 24% until March 26th, which equated to roughly 134 billion USD in market value. Solid earnings reports led to the stock price rallying and regaining the value lost by the end of May. Still, 2018 would turn out to continue being tough for the company, as it had to admit its user growth had slowed after the CA scandal. Coincidentally, the European General Data Protection Regulation (GDPR) (see Chapter 4.3) came into force around the same time. Overall, on December 31, the stock price had dropped to 131.09 USD, a roughly 35% decrease from its previous maximum of 204.87 USD. Facebook would eventually be able to leave those lows behind it, but overall this volatility and its sharp stock drops, especially the one in July after disappointing user numbers, were the biggest market value collapse ever recorded, breaking the previous record from 2000.

Overall, Facebook ended up under heightened scrutiny by authorities and privacy mistrust by its users, though Facebook would be able to overcome its losses and continue its international expansion and monetization of its services [Aie18].

### 2.5.2 Obama for America App

Some conservative public figures, such as Meghan McCain, daughter of then-Senator McCain and Republican candidate for the US presidency in 2008, claimed equality between President Obama's 2012 reelection campaign's data acquisition and targeting methods and those uncovered to be used by CA and the campaigns it provided [Tob18]. Factchecking organization PolitiFact rated this as half-true: The "Obama for America" (OfA) app would have been able to exploit the same design flaw as CA had done with its personality quiz, as restrictions only started to be implemented in 2015. The OfA app provided options to donate to the campaign, get information about how to vote and even find houses to canvas for the Obama campaign [Tob18].

A major difference consists of the transparency the apps provided their users: Whilst the CA app claimed the collected data would be used for academic research at Cambridge University instead of for profit and political campaigning, the Obama app diverged that information to the users. Beyond that difference in transparency, there are similarities in the process. The Obama Campaign used the OfA app to collect profile information of their users as well as friends lists, photos and tags (linked mentions of a specific user). This information was used to determine the users' close friends. By matching the information with other - offline - sources, people who might be able to be convinced to vote Democratic but were perceived unlikely to turn out to vote were suggested as people the user should message with a pro-Obama message [Tob18].

The acquisition of online data, comparison with offline records and determination of possible voter intention swaps does show resemblance to the more escalated version used by CA, which created profiles for all users and their friends, and then used that data for direct ad targeting of users and friends (as well as similar, lookalike, profiles) instead of asking users to message their friends.

## 2.6 Beyond Political Campaigns

This thesis is focused on disinformation campaigns in a mainly political context. However, the realization that such methods exist and are actively being used to try and sway public opinions on political topics or candidates leads to a question: What else could be done with it? Economic lobbying would be an example for that. Of course, that could still be a political angle: By trying to boost political parties and candidates whose views benefit a company, that company has a vetted interest in seeing them succeed. Generally, those companies have legal ways to do so by donating campaign money or actively making statements of support.

But what if cigarette producers wanted to reverse the decline of smoker percentages in recent decades, as the risks have become widely known? Fossil fuel producers might have an interest of reversing the public acknowledgment of fossil fuel consumption as one of the leading causes of anthropomorphic climate change. Both consequences - health issues due to smoking and the release of $CO_2$ fueling climate change - can be considered scientific consensus. Legally, there are ways those companies can try to achieve this: Lobbying with lawmakers to prevent regulation and ensure support, financing aligned candidates as well as financing studies likely to help improve their public perception are often permitted and done, though not without criticism.

Of course, disinformation campaigns could be a viable alternative for these economic stakeholders to try and influence public opinion analogous to political stakeholders. As with political disinformation campaigns, bots are a viable tool to spread information, e.g. supposedly disproving climate change. An example for such bots appearing to be in use is a network of up to 70 inauthentic accounts that was active on Twitter [Sch18a]. The network's function was spreading positive content regarding homeopathy, which is a pseudoscientific alternative medicine system [Smi12]. In case of this network, they appeared to mostly link to one specific website, homoeopathie-online.info, which belonged to the German Association of Homeopathic Doctors (DZVhä). There is no conclusive evidence as to who set up the bot network and why. This is, however, an example of bots being used in an economic context by spreading positive statements about homeopathy, which remains a big market due to its limited, but persistent prevalence in many countries [RCV+17]. The possibility that the network was a private operation by a person wanting to spread belief in homeopathy is feasible as well.

Another example revolves around the wildfires in Australia in late 2019 and early 2020 of unusual extent. The wildfires were fueled by a dry heat and winds and remained a top news topic for weeks. Climate activists used this opportunity to address climate change as a source of prolonged droughts and rising temperatures, thus increasing the likelihood of such events. However, an online network of "troll" accounts quickly attempted to spread a falsely high numbers of arsonists being arrested, implying the political left were mistaking a manmade event for climate change [Kna20]. That makes this campaign semi-political, as a political aim can be seen by putting the Australian opposition into a bad light as well as undermining a natural cause that some associated with climate change[10].

A third example of such influencing is public health - specifically, discussions about vaccinations, their efficacy and safety. The medicinal consensus is that modern vaccines are safe and effective, though that opinion is not shared by everyone in the general population. "Anti-vaxxer"[11] have

[10] This is by no means an implication that climate change caused the Australian wildfires.
[11] Colloquial term for people categorically opposing vaccinations.

become a topic in recent years, specifically named when discussing declining vaccination rates and the return of diseases like the measles. This led to measures, such as in Germany, where measles vaccinations have become effectively mandatory in March 2020 with only few exceptions[12] This topic has also been targeted by bots and Russian troll accounts on Twitter, with many of them trying to put vaccinations into a negative light or sow discord [BJQ+18]. The intention is not known, though it can be speculated that the goal is to undermine the public consensus on vaccinations and effectively lower the vaccination rate in the U.S.

Those are three examples of possible attempts to transfer strictly political disinformation campaigns into a wider range of applications, with characteristics similar to the large-scale professional campaigns described above.

As an additional and ongoing example, similar accusations have also been made during the ongoing COVID-19 pandemic. The U.S. State Department claimed that Russia deliberately sowed disinformation through associated social media accounts: In the alleged disinformation, conspiracy theories about U.S. involvement in causing the pandemic are shared, as well as intentions to leverage this to wage "economic war" on China [Gle20]. These points were allegedly shared in multiple languages over Twitter, Facebook and Instagram. No public data or report supporting these claims have been shared as of June 2020 and scientific analyses for this aspect has not been published yet.
The spread of misinformation regarding COVID-19 and the potentially deadly consequences by undermining a strong public response have led Twitter to specifically ban sharing such content on its platform [GD20]. That ban included denial of "expert guidance", encouragement of fake or non-working treatments, as well as information pretending to be from authorities or experts - effectively all content able to cause harm to Twitter users. In its April 1, 2020 update, the company claimed to have removed 1100 tweets according to those guidelines and that their automated systems had flagged "more than 1.5 million accounts which were targeting discussions around COVID-19 with spammy or manipulative behaviors" [GD20]. It has not yet been declared whether (some of) those accounts had been part of a coordinated campaign or not and how many are ultimately determined to be in violation.

Facebook, too, had announced it would remove "harmful misinformation" related to COVID-19, though the announcement did not list specific examples and properties that would constitute as such [Cle20].

[12] In Germany, children are by law required to attend school, which they now only can after their positive vaccination status has been proven.

# 3 Technical Aspects

Historically, disinformation attempts have been an analogue matter. However, the digital revolution has brought with it new technical capabilities on which social media was built as a digital way to connect with people. Now it is one of the prevalent methods of communication for many people[1].

In order to understand how disinformation campaigns are conducted on social media, some technical aspects need to be looked at. This includes methods to (un-)cover online tracks via creating or breaking anonymity, synthetic creation of fake media (deepfakes) such as altered images for propaganda purposes, or countermeasures to be employed by social media companies - which is heavily tied to the emerging field of artificial intelligence.

## 3.1 Anonymization

As mentioned in Chapter 2.2.2, one of the foundations of hiding information campaigns is disguising the fact that it is even happening. To achieve that, one of the key foundations is avoiding exposing the true origin of posts. If, for example, a single IP address from Russia were posting 1000 posts on a social network a day, even more suspiciously, with its profile info claiming the user(s) being, for example, located in the United States, detection would be easy. There are multiple ways to achieve this, with a popular method being Virtual Private Networks (VPNs), which features a secure tunnel between one client and one VPN server at a time. Another popular approach is the Tor Network, which implements a decentralized network, through which internet traffic is routed to provide anonymous browsing.

### 3.1.1 VPN

VPNs are, in essence, an extension of private networks via a public network, effectively yielding the possibility of transmitting data between two devices as if they were in the same private network [FH98]. Practically, this technology is often used to

- Access remote data securely via the internet
  - Remote connections to workplaces, e.g. work at home via remote desktop
- Use the remote device as a relay for internet connections
  - Circumvent local content restrictions e.g. imposed by ISP

[1] This is especially true in situations like the ongoing COVID-19 pandemic, when it might even be the primary or sole way of communication due to social distancing.

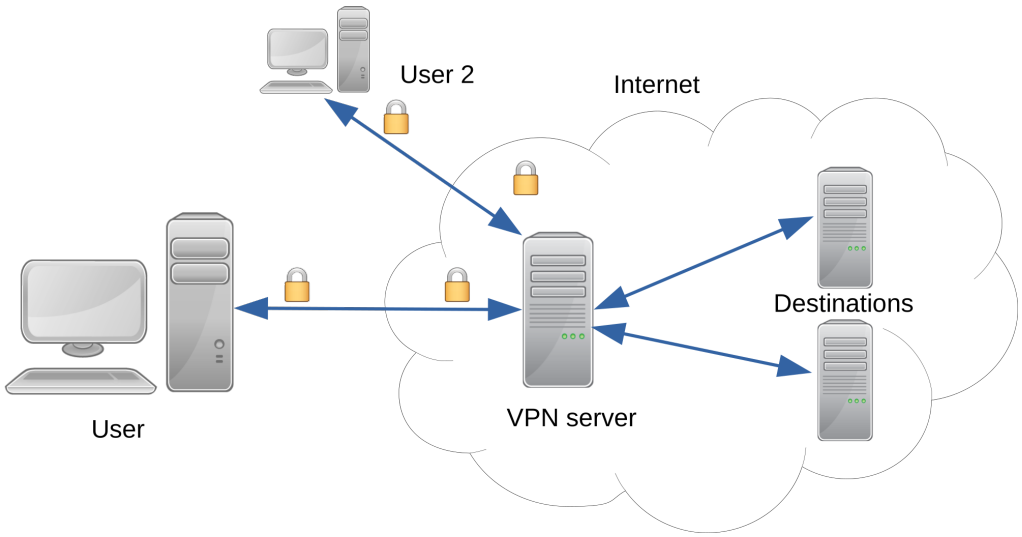


**Figure 3.1:** Simplified illustration of VPN functionality, with two-sided arrows indicating data connections, and additional locks denoting the connection to be secured by encryption.

- Avoid geo-blocking by replacing problematic IP address with another one
- Anonymously browse or otherwise connect to servers

For this thesis, only the last use-case is of relevance. Figure 3.1 provides a high-level description of how a VPN connection helps achieve this. To create a VPN connection, encrypted point-to-point (P2P) connections are created between a client and a (VPN) server, denoted in the diagram as arrows with additional locks to denote encryption. Take, for example, User 1. User 1 has established a connection to the VPN server and uses it to access one or more destinations on the internet. There are, however, more users connected and accessing other destinations. As data between users and the servers are encrypted, an outside observer cannot outright see which user accesses what destination. This can be used, for example, by people living under oppressive regimes to use the internet freely without the fear of repercussions by having their internet activity tracked (though there are methods to circumvent this). Of course, there remains a weakness in the design: If the VPN server is hacked or otherwise compromised, it is possible to expose all its users.

### 3.1.2 Proxies

Proxy servers, on the surface, display some similarities to VPNs and should therefore be mentioned when talking about the latter. Proxies, too, act as a middle-man for internet traffic - forwarding a client's requests and receiving their traffic, thus appearing as the actual sender and recipient [LA94].

The main difference is that VPN connection creates a tunnel through which all traffic from the client is routed through, always encrypted. Compared to this, a proxy is usually set on application basis (generally set as an option in that application's settings). VPN connections technically do not have to be encrypted, even though this is practically not used because the trust and security of the virtual private network practically demands this for traffic passing through the internet.

The push for secure network traffic[2] mitigates this, as the end-to-end encryption employed generally does not allow the proxy to read the content[3]. For many purposes, especially web-browsing, proxy servers can be reasonable solutions to anonymize traffic to a certain degree. They are therefore also be used for the conduction of online (dis-)information campaigns, especially if trusted proxy servers are available.

### 3.1.3 Tor

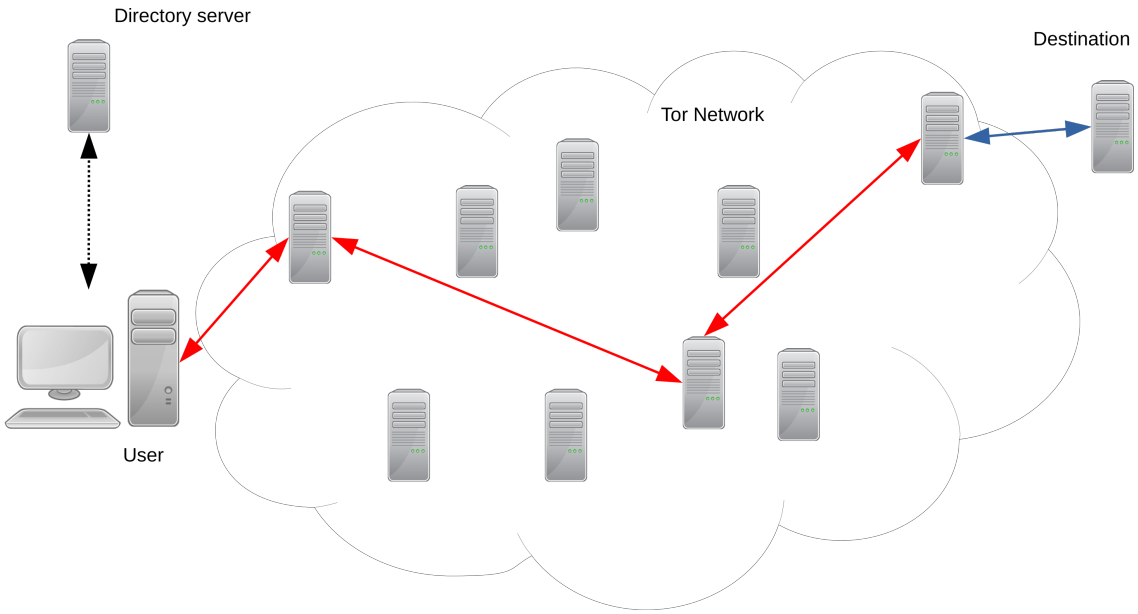


**Figure 3.2:** Illustration of Tor connection. The cloud and computers inside it illustrate the wider Tor network, with the red arrows denoting the current Tor circuit. Also present outside the network is the directory server and the actual connection destination.

Tor shares some similarities with VPN connections, in a sense that it can enable users to anonymously establish connections over a public network. A directory server provides users with a list of Tor nodes: Devices in the internet which can be used for relaying internet traffic. The basic principle relies on routing traffic between source and destinations through multiple such nodes to obscure the identity of sender and recipient [DMS04]. The distinction between different types of nodes becomes relevant when discussing possible attack vectors and measures against them, as well as legal vulnerability of node providers.

For entry points, specific nodes called "guard nodes" are used. To understand the reasoning, consider the following example: If you wanted to compromise a Tor user, you could sign up 10 devices to the Tor network. If the user's client simply chose a random subset of nodes to establish a circuit, eventually all selected nodes will happen to be by the compromising party, as a product of low

[2] By 2020, the percentage of pages loaded over HTTPS in Chrome within the United States surpassed 90%, as per Google's Transparency Report available under `https://transparencyreport.google.com/https`

[3] There are exceptions, such as companies that configure clients so the outgoing proxy can decrypt, analyze (and filter), as well as re-encrypt traffic, which requires the proxy server's certificate to be trusted by the client.

individual likelihood but an ever increasing number of attempts to hit such as subset over time. This way, any target using the Tor network would eventually be compromised. This, however, can be mitigated.

- Change the entry guard(s) less frequently
- Vet potential guard nodes for reliability and trustworthiness before allowing them as guards

This way, trying to target and compromise a specific user gets much harder, as there are much less opportunities to do so in the same timeframe. As for vetting potential guard nodes, common criteria involve sufficient uptime (avoids the possibility of quickly bringing up many guards for a specific attack) as well as connection speed and stability.

In order to establish a connection via Tor, as depicted in its most basic form in Figure 3.2 the user's device will establish an encrypted connection to an entry node, from where the data is relayed to a relay node, which in turn relays the data to an exit node. The combination of entry, relay and exit node is called a Tor circuit. The exit node then exchanges data with the original destination. Getting the response back to the user works the same way, but in reverse order. As the transmission is encrypted when passed on to the entry node, and with each entry, relay and exit node encrypting a layer with the information of where to send the package next, the general method of this way of routing is called onion routing. This concept of onion routing reaches back into the 1990s to the question of how transmission between internet users could be anonymized even if the network was being surveilled [RSG98]. The Tor network itself was initially launched in 2002, with its source code being open for transparency, though its growth and participation relied on multiple initiatives that made the usage easier for non-experts. This included the inception of easier to understand software up to the point of the development of the Tor Browser, which has the establishment of Tor connections built in by default.

A big usage case for Tor is anonymous internet usage for people who live under constant threat of surveillance and persecution in Authoritarian regimes. As such, they use it to access information sources without the threat of being monitored and punished for doing so, e.g. in countries implementing a wide censorship. The People's Republic of China (PRC) is such an example, as the state enforces strict information access rules via the "Great Firewall of China". The Tor network can be used to circumvent these restrictions, as users would not directly access prohibited hosts. However, the transparency of the Tor network is also known to regimes trying to prevent this: Due to this, directory servers can be used to extract lists of nodes and compile those into a block list, which is one of several mechanisms used to limit Tor accessibility [AF07]. This would give regimes the possibility to essentially block access to the Tor network (and flag users who try to do so). As a response, the "Tor Project: Bridges" was incepted. The project introduces bridge relays that are not publicly listed and as such not as easy to block. They can therefore be used as a bridge between the user and the Tor network, bridging traffic between the two without detection. The Bridges project has an online interface enabling users to request and see such bridge addresses (an e-mail address exists also).
Despite such measures, countries such as Venezuela in 2018 have been able to largely block access to the Tor network in times of increased protest against the government [Bra18]. The PRC is also using its internet censorship measures to try and block Tor access, including attempts to scan for such bridge nodes [WL12].

There are also Onion Services, available through the Tor Network. Originally introduced in 2004, onion services allow web services to be hidden in the Tor network, only being accessible through it [4]. This is an evolution as it establishes a service where not only the users are masked and the server publicly known, but the service provider itself may remain hidden as well. Onion services are accessed by a .onion top level domain address, just like a regular web address. The principle is used for a variety of purposes, in the media however often reduced to crime, e.g. drug markets. Onion services also allow accessing services that Authoritarian regimes outlaw, such as Facebook. It is even used by intelligence agencies such as the CIA to allow for secure information drops by informants in the field [5]. The relevance of onion services to social media disinformation campaigns is, however, limited. Those may be used to access social networks wherever such access is provided, but there is no apparent benefit beyond their already existing method of using trusted VPN or Tor access through a regular (anti-tracking optimized) web browser.

## 3.2 Deanonymization

Technologies such as the aforementioned VPN and Tor network can mask the IP address of disinformation agents overwhelmingly reliably, though not without attack vectors. However, even with successfully masked IPs, identification and reversal of anonymization is still possible, as IP addresses are not the only identifying information present.

A long-known identification method is the use of browser cookies (HTTP cookies). They consist of small fragments of information stored by the browser at the request of a website, usually in a database the browser establishes for this purpose. For example, this might be useful to let users stay logged in or identify a shopping cart on a website. Further uses include tracking for advertising. This is well known and can be relatively easily mitigated by browser addons or settings. Therefore, the IRA should be able to easily counter cookie tracking of its agents as well. However, there exist other approaches, which are harder to counter.

[4] This can be circumvented by using services as Tor2web which provides access through a gateway node. This is considered insecure, as the gateway may collect user data or may provide false content to the user as a kind of a man-in-the-middle attack.

[5] The CIA may be reached under `ciadotgov4sjwlzihbbgxnqg3xiyrg7so2r2o3lt5wz5ypk4sxyjstad.onion`

### 3.2.1 Super Cookies

Like regular (HTTP) cookies, super cookies (also called zombie cookies) are fragments of information on the user's system, intended to store information. However, unlike regular cookies, they are created to be intentionally hard to remove or remain hidden, even after applying measures intended to remove all identifying stored information, such as regular cookies [Sör13]. Samy Kamkar developed a JavaScript application in 2010, called Evercookie[6], that creates such super cookies[Veg10]. Evercookie does this by creating cookies through as many available storing mechanisms as possible, for including:

- HTTP Cookies,
- Adobe Flash (Local Shared Objects, LSO) / Microsoft Silverlight (Isolated storage),
- Hidden cookies in browser web cache,
- HTML5 storage mechanisms,
- Java Applets[7].

While some of these mechanisms have lost their significance due to the technology being deprecated (Flash and Silverlight), new mechanisms have been proposed. One such example is tracking via TLS (Transport Layer Security) session resumption. TLS is a protocol that provides secure communication over a computer network via encryption. Part of the protocol is a handshake to establish a session so that encrypted communication can be exchanged. To speed up repeated sessions, that handshake can be shortened by utilizing already exchanged key material form an earlier TLS session. This can potentially be used for user tracking not only during the short resumption span, but extended beyond, for some targets even permanently [SBFF18].
Evercookie's super cookies are especially hazardous, as the removal of super cookie data from one of the available storage mechanisms leads to Evercookie recreating it, making complete removal virus-like hard. This could potentially even happen cross-browser, if a suitable mechanism is available[8]. In a June 2012 dated presentation leaked by Edward Snowden, the National Security Agency (NSA) also names Evercookie in regard to persistent cookies that could survive common cleaning mechanisms [NSA12].

### 3.2.2 Fingerprinting

A different approach to track and identify a person online is called browser fingerprinting. Let us assume a setup such as the IRA is said to have operated under (offices with people using a computer to conduct their operations on their various accounts, see Chapter 2.2.2). An agent would have a list of daily objectives and roam social media with the accounts to be used for that purpose. Browsers, however, expose a multitude of potentially identifying information for different reasons. Those might include (not exhaustive):

[6] The code is available on Github: `https://github.com/samyk/evercookie`
[7] An example is available on Github: `https://github.com/gabrielbauman/evercookie-applet`
[8] Evercookie's Github page names Flash LSOs, Silverlight or Java as suitable mechanisms for cross-browser propagation.

- Browser addons - Extensions of functionality added by the user, for example for blocking advertisements,
- System fonts - A browser might expose this to signal the website which fonts can be displayed by being installed on the user's device,
- User agent - Part of the HTTP request to the server, commonly denotes the operating system and exact browser used (especially historically used for non-standard behavior exhibited by browsers such as Internet Explorer),
- Hardware properties - For features such as WebGL, name and features of the graphics device might be leaked, and media devices might be exposed (e.g. a listing of cameras, audio devices and microphones).

Of course, not all of those are usually unique. However, the big amount of exposed properties eventually leads to the creation of a unique combination of non-unique properties [GLB18, p. 314]. Some studies determined up to 80% of site visitors to have a unique fingerprint, later studies found lower estimates such as 30%, with a possible lowering being explained by browser techniques put in place to reduce identifying properties [GLB18]. It was also found that mobile devices tend to have lower chance of having a unique fingerprint [9]. At the same time, fingerprints not being unique can be quite unstable, with a single property changing often leading to exactly that. Therefore, fingerprinting remains a feasible deanonymization technique for browser users. Circling back to the example of IRA social media activity, recognizing the identical fingerprint on multiple accounts could raise a red flag and lead to a closer investigation, exposing clusters of accounts operated by the same device. As such, IRA-style disinformation campaigns could be susceptible to fingerprinting. Ultimately, it needs to be considered that in this IRA scenario, unique fingerprints would not be observed, as one agent operating multiple accounts would inherently make that device's fingerprint non-unique, as the fingerprint would appear on multiple accounts that, without further context, could just as well be operated by different devices showing the same fingerprint. Still, a red flag mechanism would remain, as an identified IRA account could be used to flag devices with the same fingerprint, if rare enough, for manual review. It is therefore one of many elements to unveil networks of IRA accounts.

Of course, there are measures users (and therefore also the IRA, to circle back to disinformation campaigns) can use to try and avoid the use of rare fingerprints to establish connections between accounts. Besides user addons or settings that prevent the exposure of those properties to the server, there are solutions like Tor Browser (which also implements connectivity via the Tor network by default, as the name implies) that make a privacy suite the default [10].

[9] Mobile devices tend to have a lower hardware variance (only pre-built models available, generally), less browser addons, and in general less customization than desktop devices - therefore leading to fewer rare exposed properties.

[10] As a little experiment, the device this thesis was written on was tested via Panopticlick: Firefox (default browser) with personal configuration showed a rare, but not unique fingerprint (1 in 86,292.5), default Chrome was unique (1 in all 172,759 of the last 45 days), default Tor Browser built on Firefox was much less identifiable (1 in 1,175.84).

### 3.2.3 Traffic Analysis

On December 16, 2013, multiple e-mail addresses at Harvard University received bomb threats - a witnessless crime. Analysis of the e-mails yielded the information that the person responsible used the Tor network to access Guerrilla Mail[11]. On the same day, the FBI questioned a suspect, who admitted to falsely having written the e-mails in hopes of delaying a final exam he had to face. But how could this happen so quickly, considering the perpetrator used reasonable security measures?

The answer lies within the context: Bomb threats sent to schools and colleges are often done by students of the same institution with a personal motive such as social, family or mental health problems [SSL00, p. 6]. The investigation included analysis of Harvard's network traffic, which yielded the answer: A student used the Tor network over Harvard's network at the time of the crime. By being the only one having done so at the time, he was a strong suspect immediately - and confessed [Dal13]. This simple example is counter-intuitive to the expectation, as the perpetrator's security measures were not compromised but the assumption of a student perpetrator paired with simple traffic analysis significantly weakened his anonymity simply because he uniquely stood out with his anonymization method[12].

One of the factors leading to Tor's suitability for anonymous browsing is its low latency: Data packages are not held back but forwarded immediately (see 3.1.3 for reference). This is needed for proper web browsing, as usability depends on reasonable loading times without artificial latency added to the process. Of course, this property opens Tor traffic up to pattern analysis, which can significantly weaken the anonymity even without Tor itself being broken. For example, the path of a Tor connection can be revealed through a probing attack, which doesn't reveal the source of a transmission but, by revealing the Tor nodes involved, lowers the level of security to that of a series of proxy servers [MD05]. If Tor, on the other hand, had non-trivial artificial delays implemented, correlating traffic would become harder[13]. Deanonymization remains an issue either by classical relay adversaries infiltrating Tor with compromised relays (see 3.1.3 for mitigation) or network traffic analysis [JWJ+13, p. 343]. The likelihood of compromise via network analysis was shown to be dependent on the protocol being used and the type of adversary, which could conduct traffic analysis by autonomous system (AS), Internet Exchange Point (IXP) connecting those AS, or organizations administrating multiple such IXPs[14]. Depending on the exact circumstances, some users can face a deanonymization probability of over 90% via network traffic analysis of a single IXP or AS [JWJ+13, p. 345, 347], whereas a realistic adversary contributing 100 Mbit/s via relays can reach 50% probability of deanonymizing a target in 3 months [JWJ+13, p. 342].

While trying to break or reduce the strength of anonymity is a common goal with such traffic analysis, other characteristics can be inferred with this method. By observing the rhythm and frequency of packages, the type of protocol might be inferred. For example, Secure Shell (SSH) is an encrypted network protocol to remotely access network services in a secure way - often used to remotely access systems. In interactive mode, every keystroke a user types is transmitted to the remote machine in a

[11] Guerrilla Mail is a disposable e-mail address provider. When sending e-mails through it, the sender's IP address is attached in the x-originating-ip header, which is why for anonymity, it is often used through the Tor network.

[12] In a sense, this is another version of fingerprinting, though by contextualizing and analyzing traffic patterns.

[13] In case of a single Tor user within a network, as demonstrated with the Harvard example, a reasonable delay would not have been enough to prevent correlation.

[14] The ASes connected through all IXPs comprise the internet.

separate packet immediately after pressing the key. This has been shown to be a potential weakness as the timing of those keystrokes can reveal information about length and individual keys pressed by statistical analysis, significantly weakening passwords transmitted this way if precautions are not taken [SWT01]. There is no apparent use-case for traffic analysis to break Tor anonymity in context of unmasking agents conducting a disinformation campaign on social media. The cost outweighs the benefit in this case, though this attack vector against Tor is relevant in many other constellations. Such examples include dissidents facing an adversary interested in unmasking and cracking down on dissidents, law enforcement trying to unmask felons, or intelligence agencies.

## 3.3 Deepfakes

Deepfakes are artificially created media obtained by editing video, audio or an image in such a way as to present something in a way it did not occur, typically by leveraging deep learning[15] tools [Ver20]. Typically, this will swap the person appearing in the content with another or change what the person is doing - often in form of editing voice and lip synchronization. The term itself is a portmanteau of the words deep - from deep learning, a group of machine learning methods, based on neural networks - and fake. A mostly innocuous application is in apps for "faceswaps", which has either one person's face montaged onto another or two or more people's faces swapped. Those apps are available for modern smartphones and they possess limited efficacy which usually makes the manipulation spottable.

Proof of concept also exists in political context: A popular example in Wall Street Journal's Moving Upstream has former President of the United States Barack Obama saying lines actually said by actor Jordan Peele, imitating Obama's voice, with Obama's lips artificially synchronized, changing the source footage. The result possesses reasonably close resemblance to authentic material. The impact for this in disinformation cases is obvious: By having politicians say outrageous or otherwise sensational statements, public opinion can be swayed as they may not fact-check the information, especially if it fits with a pre-existing bias, which often can significantly influence the perception of truth [BCK+19].

One approach to detecting lip-sync deepfakes is modeling facial expression and movements in order to recognize irregularities. As it often happens with new techniques, the fields of creating and detecting fakes is a cat-and-mouse game. For example, early synthesized deepfakes would either have no blinking or people blinking at unnatural times and intervals as those depictions were not included in the data used for the synthesis process [AFG+19, p. 38]. Later approaches rectified this obvious telltale sign.

### 3.3.1 Faceswaps

However, even with those more sophisticated deepfake productions, there are still intrinsic errors that can be exposed. Let us take the example of a face swap, with the person in the source material having their face replaced with that of another person. The creation usually works in the following way, in simplified description: Face detection and (2D) landmarks detection is executed on the

[15] Deep learning is a group of machine learning methods, based on artificial neural networks.

original picture, and the result being cropped. The cropped image is then warp-transformed into a standardized face (face alignment). This prepared extract is now fed into the neural network, which creates the synthesized face in the same, aligned configuration. The transformation is now reversed using the inverse matrix, and the synthetic face fused into the original face with post-processing steps obscuring potential tell-tale signs at the border areas. At this point, the deepfake is finished. [YLL19, p.1-2]

However, the fake is not perfect and an approach to detect this manipulation can work as follows, as demonstrated by Yang et al. (2019): To test for this face swap manipulation, facial landmark detection is run on the subject image. They estimate head poses by utilizing multiple landmarks either in the entire face or just in the central region. If the projected head alignment differs - as it often does, due to facial structure of two people often being different, for example - it can be a sign of patches of two people being present. If an image is genuine, such differences will vice versa not be present in such a scale [YLL19, p.2]. This, of course, can be applied to video as well, by investigating the frames making up the video.

### 3.3.2 Synthesized Speech

While much of the focus regarding deepfakes has been on visual content, audio has been a topic of interest as well. Synthetic speech creation and exposure has been in a cat-and-mouse game somewhat similar to visual content. Those synthetic voices have gained popular use cases in the last decades. Popular examples include navigation systems or voice assistants, available in a vast majority of mobile devices on the market today. However, a similar abuse akin to visual manipulation exists. This can be in conjunction with synthesized visuals (for example, a video of a politician saying something they never said - with a very close approximation of their voice and visual footage of them saying it) or independently, for example as a faked phone call or hot microphone recording. Real examples of this being used in criminal activity have already surfaced, with scams conducted through the phone. A high-profile case, by some thought of as the first AI-powered cybercrime, targeted a British energy company. In this case, the CEO of the parent company was impersonated on the phone, ordering the urgent transfer of funds in size of 220,000 EUR to a supplier. The perpetrators tried to muddy their tracks by calling back later and stating a reimbursement had been transferred by the parent company. At the third call, trying to order yet another transfer of funds, the scheme failed to do further damage as the alleged reimbursement had not arrived yet [Stu19]. This massive fraud, however, shows the dangers of this technology, especially in absence of two factor (or more) authentication.

## 3.4 Social Bots

Internet bots are, in a general sense, applications that automatically perform tasks over the internet according to their programming. These bots can be used on social media and are then called social bots. For the use in social networks, bots would typically act through its respective application programming interface (API). Such an API exposes access to multiple aspects, e.g. logging into an account with the corresponding credentials, accessing the user's feed, posting content or making search queries. This can be used for a variety of uses, such as creating third-party apps to use

social networks or for research purposes. For Twitter, for example, the API can be used to retrieve tweets by keywords for analysis or real-time retrieval by filters - which can both be useful for analysis.

To use this provided access, the API is implemented by a software library which then gives programmers access to functions to call in their code to interact with the respective social network [16]. While this covers the actual communication with the social network, a bot needs its actual functionality to be programmed as well. For a Twitter bot, this might include liking and retweeting tweets supportive of a politician, for example Donald Trump. Positive meaning could be either inferred from proper sentiment analysis or stance-qualifying keywords as well as hashtags[17]. This would then suggest public support and positive resonance for those statements as well as increase standing in the algorithms sorting and filtering contents, which factor in popularity. Other possibilities revolve around replying supportive statements in reply automatically, such as "I agree" or "Best president ever! #MAGA". This, too, can suggest higher public agreement for users and algorithms.

A more complex task is the automatic creation of content, which might be why the IRA has trollbot farms writing such content manually (see Chapter 2.2.2). One of the easier approaches to automize this might be creating a database of phrases to tweet and letting the bot randomly tweet one every now and then, possibly feeding this database on the go by adding fitting statements the bot encounters during like or retweets. True synthesis of content is much more complex. However, it is possible. One such example is Microsoft's AI chatbot Tay, which was activated on Twitter on March 23, 2016. It was soon shut down due to posting offensive content after being deliberately being repeatedly confronted with it by users, within hours. The AI, however, was able to reasonably well emulate conversations with users [NN16]. It is therefore feasible to create an AI that seeks out political content and engages in discussion with users based on pre-defined talking points, references and views. Be it to amplify or disseminate messages, the demobilization of opponents, spread of pro-government messages and political account number padding[18] has been well documented in multiple countries. Future developments, especially towards autonomous trollbots, will need to be closely monitored, as advances in making these bots act more natural will make it more difficult for users to distinguish them from genuine, human users [FVD+16].

There are, for example, various ready-to-use projects implementing Twitter bots with different use-cases and feature levels. While some automated functionality can be leveraged with online platforms such as IFTT ("If This Then That"), customizable existing codebases can be found aplenty online[19]. Conversational AI frameworks such as Microsoft Bot Framework can be used for creating "enterprise-grade conversational AI experiences"[20], for which a Twitter adapter is also available at Github[21].

[16] This design paradigm is not mandatory, though popular APIs typically have open source libraries implementing them in a variety of programming languages. This makes not using those burdensome in creating and maintaining that code when it is not required.

[17] One such example could be #MAGA, short for "Make America Great Again" - Donald Trump's 2016 slogan and rallying cry of his supporters.

[18] Automating an account to, for example, periodically follow and unfollow accounts, excessively like or retweet tweets can be used to acquire followers, as showcased in Chapter 2.2.2 under the term follower fishing.

[19] for example pyTweetBot, a Python implementation available on Github at https://github.com/nschaetti/pyTweetBot

[20] Available at https://dev.botframework.com/

[21] Available at https://github.com/BotBuilderCommunity/botbuilder-community-js/tree/master/libraries/botbuilder-adapter-twitter

## 3.5 Countermeasures by Social Networks

As social media has become a new medium to conduct disinformation campaigns, the question whether social media companies need to implement countermeasures - and how - has arisen. While lawmakers can put regulations in place, e.g. for social media companies to disclose suspicious activities or in theory even mandate active monitoring for them, the intrinsic motivation is another factor. If a social network is known to be the platform on which fake accounts are roaming around freely, and interacting with people under false pretenses, this could erode trust in the network itself, leading users to leave and ad revenues to fall.

In 2017, the U.S. Congress started investigations into disinformation campaigns on social media related to the 2016 elections, as described in Chapter 2.2.1. As a part of that, senior executives of Twitter, Google and Facebook were summoned to testify in Congress before the House and Senate Intelligence Committees as well as the Senate's Subcommittee on Crime & Terrorism. This put those companies and their handling of such infiltration into the spotlight, ultimately leading to these tech companies implementing (some) measures to address the concerns.

### 3.5.1 Twitter

As part of the Congressional hearings, Twitter committed to further cooperation and investigation into disinformation campaigns conducted on their platform. As a result of this process, the company launched the Election Integrity Hub.

As a part of this investigation, the IRA accounts mentioned in Chapter 2.2.5 were identified. Twitter notified over 670,000 users who interacted with said accounts during that timeframe. Accounts identified by the platform were not only IRA-controlled but also involved further "automated" accounts based within Russia. In its January 2018 update, the company named some additional measures with which it attempts to identify suspicious accounts due for an additional verification step. Methods for recognition involve:

- Monitoring for (near) instantaneous, unnaturally fast replies to tweets
- Detecting non-random patterns in Tweet times (e.g. bursts with no activity in between)
- Looking for coordinated engagement, as conducted by the IRA to have agents boost the content of one another

They also noted changes in verifying accounts, as in phone number verification and reCAPTCHAs for detecting activity of bots rather than humans [Twi18].

Interestingly, further plans for improvements within 2018 specifically named increased utilization of machine learning to detect fake accounts and recognize bots. Whether Twitter implemented such machine learning measures – and if they did, how – does not appear to be publicly known. In June 2019, Twitter acquired the machine learning startup Fabula AI from London. In its blog post announcing the acquisition, Twitter's CTO Parag Agrawal stated the company would be joining Twitter's own "Cortex" team focused on machine learning applications for its platform [Agr19].

Fabula AI's focus led on graph deep learning. In context for social media, such a graph could be comprised of users being the nodes and the edges between them the interactions between them. The company's focus lied within detection of manipulation within a network, and the special big

data expertise required to handle those huge data sets which a social network of this size would easily produce. This can be used for detection of anomalous activity and therefore detection of (dis-)information campaigns. Twitter also implemented changes to make conducting campaigns with automation via the Twitter API harder, e.g. by making it more difficult to operate with multiple accounts on power user tools such as TweetDeck[22].

Twitter also addressed the issue from another perspective: Introducing verification steps to make disinformation campaigns as had been seen in 2016 more difficult. In its revised guidelines introduced in 2018, Twitter established the need to get certified, with different requirements depending on the jurisdiction the election was happening in[23]. For example, for the European Parliament elections in 2019, individuals needed to provide a EU member state-issued photo ID, whereas candidates were mandated to submit proof of their candidacy registration, and miscellaneous organizations needed to provided their EU VAT or company identification number.
Candidates for an election had their tweets marked with a badge marking them as a political candidate for that election, and political ads were mandated to be connected to a Twitter profile under certain identification requirements and representing a political campaign. After applying this system for the U.S. midterm elections of 2018 and said EU parliamentary elections of 2019, Twitter announced that, in contrast to Facebook, it would ban all political ads entirely [Fei19b].

### 3.5.2 Facebook

Unlike Twitter (and Google), Facebook's response to the disinformation campaign(s) executed on its platforms Facebook and Instagram has been much less drastic. In the run-up to the 2020 election, Facebook declined to limit the capability for micro-targeting via specific demographic and interest-based parameters, and refused to disallow political ads containing false information. Facebook's CEO Mark Zuckerberg stated that "political speech is important" and Facebook did not want to interfere with that [OA20].

However, the point that disinformation could be legally and openly advertised on Facebook faced criticism. During a hearing in congress, for example, U.S. Democratic Representative Alexandria Ocasio-Cortez asked Zuckerberg if she were able to run advertisements claiming Republican Senator Lindsey Graham supported the Green New Deal, a large proposed bill addressing climate change and income equality [Cul19]. That bill has been faced by strong opposition by Republicans and Graham does not support it. However, after that hearing, a Political Action Committee (PAC) published an ad on Facebook which stated exactly this to test Facebook's willpower to maintain that point. The ad, however, was removed as it was not entered by a (U.S.) politician but a PAC and thus eligible for review, which it failed [Cul19]. Facebook still retains this position.

Overall, Facebook is following the path of searching for disinformation campaigns on its platforms and publishing their findings which it started during cooperation with Congressional investigations into the matter. For example, as recently as in December 2019 Facebook suspended a network of over 100 fake accounts that had been used to "seed false narratives online targeting Ukraine and

[22] TweetDeck is a free in-browser dashboard for Twitter that enables users to simultaneously use multiple accounts.

[23] While superseded now, the guidelines are still online at `https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content/how-to-get-certified.html`

other countries in Eastern Europe" [Stu20]. As the CA scandal has shown (Chapter 2.5.1), the amount of trust is a factor in its stock pricing and therefore the appearance of its content as genuine in its own interest.

## 3.6 Automated Detection of Disinformation

In the preceding sections, some approaches to tackling disinformation campaigns have already been introduced. One of those basic measures is fact-checking. To make this a feasible approach, posts need to be tagged for verification automatically, e.g. by recognizing strong traction, polarized responses or automated behavior acting upon such content, to then be checked by people and afterwards being marked as true, false or contested (or in a similar scale).
Take, for example, rumor analysis, which can be performed in four phases [AGH19, p. 32]:

1. **Detection** - This can be done by the appliance of Artificial Intelligence to identify rumors as they surface.
2. **Tracking** - Once a rumor has been detected, posts need to be classified as being thematically related. This could be done by picking up certain keywords or hashtags as topical identifiers.
3. **Stance Classification** - Categorizing thematically related posts by their stance on a topic, for example. Stances include classification of truthfulness (true / false), querying ("Is it true that pineapple is internationally rated as the worst pizza topping?") or commenting ("Interesting rumor about pineapple pizza today!").
4. **Veracity** - While AI judgment is in development with limited success, this step usually involves humans with expert knowledge, as this is often a process that involves complex decisions. Depending on the fact checking context, reasonings for judgments are also provided as supplemental information.

However, not all information is as machine-readable as simple texts posts. Rumors - or in case of intentionally false rumors, disinformation - can also be put into the form of memes[24] to transfer a message in a more subtle way. Further research will need to be conducted to fully automize the process and make more content formats available for automation.

Attempting to recognize - and tag or even remove - false information as it is posted, is only one of multiple approaches. For example, distributing a false rumor is not necessarily disinformation if it is done in good faith. This procedure alone might help users recognize good from bad information but does not qualify the intent.

Recognizing the artificial amplification (or creation) of rumors, for example, can give hints towards intent. In that context, artificial amplification might consist of automated behavior as well as fake accounts controlled by humans. Going back to the example of the IRA, the objective could be to amplify negative rumors about a political candidate. The rumor might, but does not have to be placed by the IRA in the first place. However, by making sure the rumor gets exposure, it

[24] A meme is "an amusing or interesting item [...] spread widely online especially through social media ", per Merriam-Webster.

will get uninvolved people to take note and start discussing it, therefore kickstarting the spread of misinformation. Social networks generally do not allow automated behavior for benign or malicious reasons, and therefore try to recognize this as it occurs.


**Figure 3.3:** Illustration of the Agent, Message, Interpreter conceptual framework, with the smaller elements on the right denoting attributes and possible values. Image source: [WD17]

An approach introduced is the Agent Message Interpreter (AMI). The AMI is a conceptual framework emphasizing not only the need to understand the message of a disinformation campaign but also the agent(s) delivering it (source and propagators), its reach and physical impact (e.g. changed voting). Therefore, distinguishment and consideration of three different aspects has been called for in order to better understand disinformation campaigns in their entirety [WD17], and as seen in Figure 3.3:

- **Agent** - Autor, propagators of the message - as well as their motivation.
- **Message** - The content itself, form of expression, target audience and its accuracy.
- **Interpreter** - The people reading the message as well as the message's effects on them (change in belief, opinion).

#### Agent

Starting with assessing the agents, checking news sources for hyper-partisanship or a bias is a first step to detect at least some agents: Typically, state-owned TV stations from authoritarian nations fall into this category, where at least the news composition is altered in a way to bolster the regime's intentions. This for example, could naively be done by sentiment analysis of certain topics (positive and negative reports e.g. about a politician) in comparison to the broad average of the free press

to detect outliers in either direction. The Global Disinformation Index is such a project building a system rating the reliability of news sources worldwide, at real-time [AGH19, p. 36].
However, this is not the only way agents might try to spread information. As covered by exploring the methods of the IRA, creating false accounts to give the impression of a message being transmitted by various independent sources - private people, news organizations, government agencies - credibility can be established, and the exposure be built up by starting the process of misinformation.

#### Message

Tackling the message aspect has already been partially touched upon above, with the example of rumor detection. Noteworthy is the discrepancy between a claim's veracity - its objective truthfulness - and its credibility - its perceived truthfulness. An example for widening this gap is the improved manipulation potential of deepfakes, which can be debunked but might be sufficiently credible to have an impact on opinion, e.g. by going viral, its content becoming ingrained in opinions while the refutation's spread might not have the same virality and impact, leading to a net-gain for the agents involved.

Of course, in order to analyze the message, there always is the possibility of manual review. Humans are arguably good at identifying intentions and core statements in man-made content, as well as identifying context and its impacts on meaning (e.g. irony). On its own, this is not viable as the manpower required to manually view such amounts of data is higher than the amount of resources typically available. While emotion can be detected automatically, analysis of credibility remains an open issue. Ongoing and future research might able to improve recognizing linguistic clues and apply linguistic analysis to rate credibility in an effective and efficient way [ZRM+18].

#### Interpreter

While knowledge of the audience and impact on it is classically derived from statistics such as content impressions or Like / Share numbers, those are inherently limited: No proper assessment of the actual impact - change in behavior or thinking - can be derived from this. While this is still a big knowledge gap, some research has been done, presented in Chapter 2.3.

The PHEME project has conducted research attempting to recognize rumor stance - a form of interpreter reaction to the message - in social media users. Machine learning algorithms are in development and can get significant accuracy (up to 80%) in a limited test scenario (effectively a proof of concept), though with room left to improve [KLA17]. Of course, in order to classify a user's stance on a rumor in real scenarios, automized detection of rumors needs to be functional in the first place. Improved understanding of message analysis is therefore beneficial for interpreter analysis.

# 4 Public Response

As already touched upon in Chapter 2.2.1 with the 2016 U.S. election, the age of social media has brought with it new challenges for lawmakers. While disinformation campaigns are certainly one concern, there are many related issues that have come into the spotlight. For examples, the emergence of deepfakes and their use for harassment or propaganda has led to initiatives by lawmakers to regulate them. At the same time, scandals involving (mis-)use of people's data led to pushes for increased transparency and data protection for users.

## 4.1 Australia

In Chapter 3.3, deepfakes were introduced and their abuse potential explored. There is one particular example that not only led to public attention but also led to new laws. However, the obvious risk the technology poses has already had real life impact, as in the example of Noelle Martin, an Australian woman. At the age of 18, she had discovered that there were pornographic images online with her in it - even though she had never participated in it. Over the years, her personal fight led her to request takedowns and speak out for legal protection, which led her down the road of becoming a well-known advocate for criminalizing nonconsensual deepfaking. Her activism has been credited for recent laws in Australia criminalizing such acts, as introduced on the federal Australian level in 2018 [Sch19].

On a much larger scale, the law changes included specifically making revenge porn illegal, which describes the nonconsensual release of pornography, often recorded during a relationship and then released after a breakup, out of "revenge".

## 4.2 California Consumer Privacy Act (CCPA)

A bill outlawing the same "revenge porn" deepfakes as in Australia has also been passed in California in 2019. Furthermore, another bill has been drafted, passed and enacted in California during the same timeframe. While the outlawing of deepfakes in Australia applied to private matters, the California bill targeted political use of deepfakes to influence an election. The bill states the intent to "prohibit a person, committee, or other entity, within 60 days of an election at which a candidate for elective office will appear on the ballot, from distributing with actual malice materially deceptive audio or visual media of the candidate with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate, unless the media includes a disclosure stating that the media has been manipulated" [Sta19].

This 2019 California bill is a unique landmark law specifically addressing the potential of deepfakes to be used as content in disinformation campaigns and remains unregulated in many jurisdictions around the globe. In 2018, a ballot initiative for GDPR-like legislation - a way of enforcing a state-wide referendum over a bill in California - gained significant traction and was able to accumulate over 600,000 support signatures. Its emergence was seen as public dissatisfaction with the current situation, encouragement by the rise of the European GDPR and, specifically, the Cambridge Analytica scandal and the lack of consequences from it [Gho18]. Over concerns of economic regression by overregulation of the state's IT sector, the California Chamber of Commerce sponsored the launch of the Committee to Protect California Jobs to publicly lobby against the advance of the ballot initiative. In fact, as Fig. 4.1 shows, multiple international but US-based IT giants such as Amazon, Facebook and Google contributed to the campaign against the proposal, with Microsoft instead giving the campaign loans (all in yellow). Furthermore, only some of the companies are seated within California. Amazon and Microsoft, for example, are based in Washington. Interestingly enough, several large telecommunication companies such as AT&T, Comcast and Verizon are on that list. Those three are also the largest companies in their field within the United States and naturally are in possession of large sets of private communications data of their customers.

| NAME OF CONTRIBUTOR | PAYMENT TYPE | AMOUNT | TRANSACTION DATE |
|---|---|---|---|
| ALLIANCE OF AUTOMOBILE MANUFACTURERS, INC. | MONETARY | $200,000.00 | 5/14/2018 |
| AMAZON.COM, INC. | MONETARY | $195,000.00 | 6/8/2018 |
| AMERICAN ASSOCIATION OF ADVERTISING AGENCIES | MONETARY | $25,000.00 | 5/21/2018 |
| ASSOCIATION OF NATIONAL ADVERTISERS, INC. | MONETARY | $50,000.00 | 5/29/2018 |
| AT&T INC. AND ITS AFFILIATES | MONETARY | $33,436.00 | 11/16/2018 |
| AT&T INC. AND ITS AFFILIATES | MONETARY | $200,000.00 | 3/7/2018 |
| CALIFORNIA NEW CAR DEALERS ASSOCIATION ISSUES PAC | MONETARY | $200,000.00 | 5/15/2018 |
| CHARTER COMMUNICATIONS | MONETARY | $25,000.00 | 8/13/2018 |
| COMCAST CORPORATION | MONETARY | $33,436.00 | 10/30/2018 |
| COMCAST CORPORATION | MONETARY | $200,000.00 | 2/14/2018 |
| COX COMMUNICATIONS, INC. | MONETARY | $50,000.00 | 6/12/2018 |
| DATA & MARKETING ASSOCIATION | MONETARY | $50,000.00 | 6/5/2018 |
| FACEBOOK, INC. | MONETARY | $200,000.00 | 2/27/2018 |
| FIRST AMERICAN FINANCIAL CORPORATION AND AFFILIATED ENTITIES | MONETARY | $10,000.00 | 3/12/2018 |
| GOOGLE, LLC | MONETARY | $200,000.00 | 2/28/2018 |
| GOOGLE, LLC | MONETARY | $33,436.00 | 9/26/2018 |
| INTERACTIVE ADVERTISING BUREAU | MONETARY | $50,000.00 | 6/12/2018 |
| MICROSOFT CORPORATION | LOAN | $195,000.00 | 6/1/2018 |
| MICROSOFT CORPORATION | LOAN | $0.00 | 6/1/2018 |
| MICROSOFT CORPORATION | LOAN | $0.00 | 6/1/2018 |
| NETWORK ADVERTISING INITIATIVE | MONETARY | $25,000.00 | 5/25/2018 |
| PERSONAL INSURANCE FEDERATION COMMITTEE | MONETARY | $50,000.00 | 6/15/2018 |
| UBER TECHNOLOGIES, INC. | MONETARY | $50,000.00 | 5/31/2018 |
| VERIZON COMMUNICATIONS INC. AND ITS AFFILIATES | MONETARY | $200,000.00 | 2/22/2018 |

**Table 4.1:** Extracted campaign finance data via the Secretary of State of California's website [Cal].

In turn, the ruling Democratic Party of California swiftly introduced and subsequently passed a compromise bill that encompassed most, but not all of the proposed changes in it due to concerns of economic downturns brought on by hindering the state's big IT sector [MB18] [Fei19a]. In Section 2 (g) of the California Consumer Privacy Act of 2018 (CCPA) bill, the CA scandal is directly cited as one of the motivations of the bill.

> In March 2018, it came to light that tens of millions of people had their personal data misused by a data mining firm called Cambridge Analytica. A series of congressional hearings highlighted that our personal information may be vulnerable to misuse when shared on the Internet. As a result, our desire for privacy controls and transparency in data practices is heightened.

The CCPA, in its provisions and goals, is similar to the European GDPR legislation passed two years prior. It establishes and underscores the following rights, with violations causing the danger of financial fines [Cal18]:

1. Knowing what personal information is collected about them.
2. Knowing if their personal information is sold or shared with third parties (and who).
3. Option to decline the sale of their personal information.
4. Accessing the personal information collected about them.
5. No disadvantages in pricing or servicing options if any of the rights above are exercised.

However, there are some differences compared to the GDPR. While the GDPR mandates the user having to opt-in for any data collection via browser cookies except those necessary for functionality, the CCPA does not contain such a provision. Also, the scope is much narrower: Only California-based companies, whose primary business model is the sale of personal information, or with a yearly revenue above 25 million USD are targeted. This does not only exclude most smaller companies; it also does not enact its protections for Californians interacting with out-of-state companies in a way the GDPR does. The opt-out right also functions somewhat differently: The CCPA mandates an option (link) on websites to opt-out of selling their personal information to third parties, whereas the GDPR has provisions for opting out of data processing for marketing purposes as well as withdrawing the consent for processing in the first place.

#### Effects

As of June 2020, no information regarding the law's impact from the state's point of view (cases of non-compliance) after coming into effect on January 1, 2020 was available. This is, in part, due to the six-month grace period until Summer 2020. However, some paradoxical privacy issues have surfaced for end-users. A company called "i360", whose business model revolves around advertising and data collection (and thus falls under the CCPA due to its data activity), requested that users, for verification purposes, submitted their full Social Security Number (SSN). This was done as a form of mandating companies to verify the submitter's identity to a "reasonable degree of certainty", though to an unnecessary specific degree. The SSN is a sensitive information and its potential leakage especially dangerous, as usually it is permanent once assigned and its misuse potential is vast. Prominently, the SSN can be used to facilitate identity theft.

Other companies have been in the spotlight for similar requests, such as requesting users to upload their Driver's License - in the U.S. commonly the primary form of identification - or, as done by Comcast, additionally to that, a selfie. Also in the public eye is the practice of the AI startup "Clearview AI", which made headlines due to creating a surveillance system and database by scraping social media networks for users' photos. Whether or not this data is deleted as soon as

the reason for requesting it in the first place - verification - was obsolete, is not clear from public information so far. Another interesting point is the creation of a new niche market segment for startups that act as a broker between companies and users for CCPA (and GDPR) requests. Those would offer users to act as an easy gateway to make such information requests, but have similar underlying privacy concerns due to how they function, e.g., with e-mail access, as is the case with the startup "Mine" [Whi20].

## 4.3 General Data Protection Regulation (GDPR)

The GDPR is a privacy regulation in the European Union (and the European Economic Area). Its purpose is to make data processing more transparent to the data subjects, make privacy the default, ensure data security and limit the scope and timeframe of collection. It was first proposed by the European Commission in 2012 and adopted in 2016, while its provisions became applicable in 2018.

The GDPR introduces the following principles, which it claims to enforce, in its provisions:

- **Lawfulness, fairness and transparency** - The processing of personal data should be transparent to the subject.
- **Purpose limitation** - Data should only be collected and used exclusively for reasonable purposes stated and agreed upon by the subject (with some exception to science, archiving and public interest).
- **Data minimization** - Only data required for the stated purposes should be collected.
- **Accuracy** - Personal data should be up to date if collected and the collector needs to undertake steps to ensure the correctness (e.g. subjects have a right to correct wrong data for credit score calculation).
- **Storage limitation** – Unanonymized data may only be stored while it is required to fulfill purposes stated (again with some exceptions, e.g., archiving).
- **Integrity and confidentiality** - The security of data against breaches, misuse or unintended deanonymization needs to be ensured by technical (e.g. encryption) or organizational measures.
- **Accountability** - The controller of data is responsible to uphold the data principles above.

It has been described as a landmark privacy regulation but also faced criticism both on not being decisive enough in some parts, and by being a burden to some businesses (e.g. by requiring many businesses to have a certified data protection officer employed or having to acquire and administrate data processing consent forms). In some cases, uncertainty on how to properly comply or implement provisions also existed, as no court precedents for unclear questions existed yet (as is common for completely new laws).

However, its provisions might also lower the risks of CA scandal-esque data abuse. For example, using data for election ad targeting secretly, after collecting data via an app, is clearly a GDPR violation and as such subject to a large fine and criminal proceedings. As the GDPR was not in effect during the time of the breach, its provisions were not applicable. Because of this, the maximum

punishment of 500,000 GBP was ruled in this case. Under the GDPR, the maximum possible fine would have consisted of 4% of annual global turnover or 20 million EUR (whichever higher). For Facebook, with 2017 turnover, that would have been up to 1.3 billion EUR - higher by a factor of about 2300 [HP18]. Therefore, privacy laws are a crucial piece in dealing with breaches simply by making it costly not to adhere to privacy standards. In turn, this would make acquisition of microtargeting data harder for agents conducting disinformation campaigns, and thus one possible way to lower their impact.

## 4.4 Germany

With the dawn of the refugee crisis[1] in 2015, political discourse in Germany has become more polarized and heated. The political landscape started to change, with the "Alternative für Deutschland" (Alternative for Germany, short AfD) becoming a gathering point for different political trends in the spectrum from right of center to right radicalism. Initially, in 2013, the party started out as a kind of national liberal and populist movement criticizing the Euro currency and related policies, as well as generalized EU criticism. However, the party started to shift towards right-wing and far right political stances with leadership changes in 2015 and 2017 [Lee18]. In an increasingly polarized political climate, the party reached the second highest voter share in three state elections in Eastern Germany in 2019.

### 4.4.1 Network Enforcement Act (NetzDG)

As a result of the Nazi dictatorship, Germany, unlike many other countries, has laws that prohibit inciting hate on protected minorities with untruthful statements. The point of prohibiting incitement to hatred (Volksverhetzung) is to prevent building a hateful stereotype as was done with the Jewish in the 1930s, first by the NSDAP and then on a broad level in society, media and the entire government.

Besides relatively strict laws on libel, slander and insults, the prosecution of Volksverhetzung has been seen as a challenge, especially as many social media companies are located in the United States, where such law does not exist. In 2017, Germany passed a new law to more effectively combat hate speech on social media platforms, called Netzwerkdurchsetzungsgesetz (Network Enforcement Act, NetzDG). The NetzDG does not introduce a specific offence of disinformation. Its aim is to increase enforceability of the German criminal code in social media, as there had been a perceived lack of accountability in networks hosted outside of Germany. For example, the United States has less strict laws regarding incitement to hatred, insults and denial of the holocaust or usage of Nazi symbols in public.

Therefore, social networks and similar platforms now must enable users to report content they believe to fall under such offences, review those reports, and remove them if they find those reports to be (likely) accurate. Failure to do so would lead to fines, which might cause social networks to be overly sensitive and prefer deleting more content than necessary rather than being fined for failing

[1] The refugee crisis was a period of years during which an increased number of refugees sought asylum in the European Union, peaking in 2015 and 2016.

to remove indeed illegal content. Similarly, there is also criticism that companies would essentially rule on legality as well as lackluster options for users to appeal their decisions, thus potentially undermining due process[TL19].

Furthermore, social networks also are obliged to provide personal information of accused offenders. NetzDG is similar to the GDPR in the sense that extends applicability to social media companies not located in Germany but have more than two million users in Germany, thus gaining options to enforce German law on companies having users in large numbers in Germany (§1 NetzDG). The law also intends that platforms, to which the law applies, have to publish regular transparency reports in which they disclose the amount of requests they received (e.g. from German law enforcement) and how often they disclosed the requested data or deleted content that had been flagged as illegal . Reception has been mixed, as polls by pollster Civey, conducted in 2018, have indicated a majority finds the legal obligation to delete illegal content on social networks positive[2] while more people believe it is reducing free speech[3].

### 4.4.2 Tagesschau Faktenfinder

As the topic of fake news became more public attention - with the term often being associated with U.S. president Donald Trump - news outlets considered methods to provide corrections to circulating false rumors, mis-, and disinformation, as well as propaganda cases. In response to this, Tagesschau - a news show by German public television - launched its online segment Faktenfinder, which it dedicated to analyze such rumors and add supplemental information as well as provide the truth in case of wrong information [Kas17].
For this, they select current topics being discussed online with a certain degree of polarization and what they see as misinformation getting traction. In their articles, they name the information in question and subsequently add context, factual correction, or explanation without trying to boil it down to labeling it as true or false [Kas17]. The segment is still being published on their website. Systematic review of its accuracy has not been published, which would be an interesting aspect for further research into fact-checking the Tagesschau fact checkers.

[2] Online available at https://civey.com/umfragen/2285/wie-beurteilen-sie-dass-betreiber-von-sozialen-netzwerken-jetzt-gesetzlich-verpflichtet-sind-rechtswidrige-inhalte-zu-loschen

[3] Online available at https://civey.com/umfragen/2287/denken-sie-dass-das-neue-netzwerkdurchsetzungsgesetz-netzdg-die-meinungsfreiheit-einschrankt

# 5 Conclusion and Outlook

This thesis gave an overview over the emergence of disinformation campaigns, before and during the age of social media, and gave an overlook over adjacent topics. First, the proper terminology was introduced, historical backgrounds explained, and the prime example of the U.S. election of 2016 analyzed. The evidence provided both by public and private sector gave a strong indication that Russia conducted a disinformation campaign to influence the outcome of that election in social networks such as Facebook, Instagram and Twitter. However, the question whether these campaigns ultimately had a significant impact on voters remains unanswered, as this turned out to be a research niche with only limited publications available. Also, Big Data was introduced and specifically the CA Scandal analyzed, with its links to recent disinformation campaigns being displayed, raising ethical and legal questions in the field of targeted advertising. Furthermore, traces for information campaigns in other contexts - such as economic lobbying - were introduced, underlining the increasing significance of online disinformation attempts. Especially in the context of dynamic and fast-evolving crises such as COVID-19, the potential for abuse seems high, indicating a need for further research.

This was followed by a look at the technical aspects, starting with anonymization techniques useful for disguising origins of fake accounts, followed by an overview over the broad topic of deepfakes, another auxiliary aspect. By creation of fake videos or audio, false evidence can be created and as such the distinction between real and fake is made more difficult. The emergence of new deep learning techniques has given conductors of disinformation campaigns a new way to be supplied with content for distribution. Together with improved insight into target audiences by application of big data, this could make disinformation campaigns much more powerful. At the same time, social media companies try to apply the same technological advance to detect these campaigns more effectively, and this will be a point of interest for further research.

The third part of this thesis introduced and described public measures around the world to address disinformation campaigns and related topics, such as big data and deepfakes. The European GDPR and Californian CCPA are described as two high-profile measures to give users more control over their data as well as strengthen privacy requirements. Meanwhile, countries like Germany have introduced public media formats such as Faktenfinder to uncover fake news amidst growing political polarization. Australia, on the other hand, is a pioneer at deepfake regulation.

Ultimately, this thesis finds that disinformation campaigns are happening in social media. While their efficacy is not yet determined, the risk - especially considering symbiotic effects due to developments in related fields such as big data and deep learning - is significant. Public scrutiny is still in an early stage, and many jurisdictions have not yet published sufficient information regarding social media disinformation campaigns, such as the United Kingdom, where the relevant Brexit report is not yet released to the public. Regulation is not a trivial task, as attempts to combat this type of disinformation are able to undermine personal privacy or even freedom of expression. There does,

therefore, not seem to be a consensus in Democratic nations as to how to keep personal freedoms and liberties intact while addressing this new form of propaganda, even if effective, all-purpose measures were to be introduced.

As mentioned in Chapter 2.2.6, the U.S. election in November 2020 will be a major test for efforts to contain disinformation campaigns in social media. As companies and governments had 4 years' time to analyze and respond to those activities, this will be a test of their measures to identify and combat campaigns rapidly. Further research should, and likely will, also investigate what impact those campaigns have in the first place. Yet, much remains to be analyzed regarding the 2016 election four years after. While, for example, Twitter archive analyses have established agents and discussed topics, this could be combined with sentiment analysis in further research. While humans can manually establish the meaning very well, machine learning's advances offer ways to predict tweets' sentiment automatically. As such, this could be used to establish timelines of when which candidates were talked positively about, when negatively, if this remained constant throughout or if there were changes in attitude. And if so, what happened during those timeframes? The impression that the question of *what* has been better solved than the *why* remains and calls for further interdisciplinary research.

This field is in fast and constant change, as new technologies for detecting and conduction of disinformation campaigns are constantly being developed. Therefore, further research and monitoring is essential to respond to developments as they happen and to gain further insight. As of 2020, the surface has only been scratched.

# Bibliography


[AF07] S. Afroz, D. Fifield. *Timeline of Tor censorship*. 2007 (cit. on p. 46).

[AFG+19] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li. "Protecting World Leaders Against Deep Fakes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 38–45 (cit. on p. 51).

[AGH19] A. Alaphilippe, A. Gizikis, C. Hanot. *European Parliament Panel for the Future of Science and Technology: Automated tackling of disinformation*. 2019. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf (cit. on pp. 56, 58).

[Agr19] P. Agrawal. *Twitter acquires Fabula AI to strengthen its machine learning expertise*. June 2019. URL: https://blog.twitter.com/en_us/topics/company/2019/Twitter-acquires-Fabula-AI.html (cit. on p. 54).

[Aie18] C. Aiello. *What data scandal? Facebook's stock notches an all time high, shrugging off user privacy woes*. July 2018. URL: https://www.cnbc.com/2018/07/06/facebook-hits-all-time-high-marking-full-recovery-from-data-scandal.html (cit. on p. 39).

[BAB+18] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky. "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221 (cit. on pp. 31, 32).

[BCK+19] M. Babaei, A. Chakraborty, J. Kulshrestha, E. M. Redmiles, M. Cha, K. P. Gummadi. "Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking." In: *FAT*. 2019, p. 139 (cit. on p. 51).

[BGM+20] C. A. Bail, B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, A. Volfovsky. "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017". In: *Proceedings of the National Academy of Sciences* 117.1 (2020), pp. 243–250 (cit. on pp. 31–33).

[Bit90] L. Bittman. "The use of disinformation by democracies". In: *International Journal of Intelligence and CounterIntelligence* 4.2 (1990), pp. 243–261. DOI: 10.1080/08850609008435142. eprint: https://doi.org/10.1080/08850609008435142. URL: https://doi.org/10.1080/08850609008435142 (cit. on p. 15).

[BJQ+18] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, M. Dredze. "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate". In: *American journal of public health* 108.10 (2018), pp. 1378–1384 (cit. on p. 41).

[BLW18] B. C. Boatwright, D. L. Linvill, P. L. Warren. "Troll factories: The internet research agency and state-sponsored agenda building". In: *Resource Centre on Media Freedom in Europe* (2018) (cit. on p. 29).

[Bog09] T. Boghardt. “Soviet Bloc Intelligence and its AIDS disinformation campaign”. In: *Studies in Intelligence* 53.4 (2009), pp. 1–24 (cit. on p. 15).

[Bra18] R. Brandom. *Venezuela is blocking access to the Tor network*. June 2018. URL: https://www.theverge.com/2018/6/25/17503680/venezuela-tor-blocked-web-censorship (cit. on p. 46).

[Cal] California Secretary of State Alex Padilla. *Campaign Finance: COMMITTEE TO PROTECT CALIFORNIA JOBS, SPONSORED BY THE CALIFORNIA CHAMBER OF COMMERCE*. URL: http://cal-access.sos.ca.gov/Campaign/Committees/Detail.aspx?id=1401518&view=received&session=2017&type=all (cit. on p. 60).

[Cal18] S. of California. *Bill Text - AB-375 Privacy: personal information: businesses.* 2018. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375 (cit. on p. 61).

[Cen86] Central Intelligence Agency (CIA). *Soviet Disinformation: Allegations of US Misdeeds (Memorandum)*. Declassified and released in 2011. Mar. 1986. URL: https://www.cia.gov/library/readingroom/docs/CIA-RDP86T01017R000100620001-1.pdf (cit. on p. 14).

[Che15] A. Chen. *The Agency*. 2015. URL: https://www.nytimes.com/2015/06/07/magazine/the-agency.html (cit. on p. 20).

[Cle20] N. Clegg. *Combating COVID-19 Misinformation Across Our Apps*. Mar. 2020. URL: https://about.fb.com/news/2020/03/combating-covid-19-misinformation/ (cit. on p. 41).

[Cra18] D. Crary. *Religion and Right-Wing Politics: How Evangelicals Reshaped Elections*. Oct. 2018. URL: https://www.nytimes.com/2018/10/28/us/religion-politics-evangelicals.html (cit. on p. 23).

[Cra19] D. Crary. *Trump steadily fulfills goals on religious right wish list*. Aug. 2019. URL: https://apnews.com/c8626c6bdbab4e3f8232ea1499a6954b (cit. on p. 21).

[Cul19] E. Culliford. *Facebook takes down false ad from PAC on Republican Graham*. Oct. 2019. URL: https://www.reuters.com/article/us-usa-election-facebook/facebook-takes-down-false-ad-from-pac-on-republican-graham-idUSKBN1X50IZ (cit. on p. 55).

[Dal13] T. M. Dalton. *AFFIDAVIT OF SPECIAL AGENT THOMAS M. DALTON*. Dec. 2013. URL: http://cdn3.sbnation.com/assets/3738299/kimeldoharvard.pdf (cit. on p. 50).

[Dav15] H. Davies. *Ted Cruz using firm that harvested data on millions of unwitting Facebook users*. 2015. URL: https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data (cit. on p. 38).

[DGG15] A. De Mauro, M. Greco, M. Grimaldi. “What is big data? A consensual definition and a review of key research topics”. In: *AIP conference proceedings*. Vol. 1644. 1. American Institute of Physics. 2015, pp. 97–104 (cit. on p. 36).

[Dha11] M. K. Dhami. *Behavioural Science Support for JTRIG’s (Joint Threat Research and Intelligence Group’s) Effects and Online HUMINT Operations*. 2011. URL: https://theintercept.com/document/2015/06/22/behavioural-science-support-jtrig/ (cit. on p. 17).

[DHS16] DHS Press Office. *Joint Statement from the Department Of Homeland Security and Office of the Director of National Intelligence on Election Security*. Oct. 2016. URL: https://www.dhs.gov/news/2016/10/07/joint-statement-department-homeland-security-and-office-director-national (cit. on p. 19).

[DI19] A. Dawson, M. Innes. "How Russia's Internet Research Agency Built its Disinformation Campaign". In: *The Political Quarterly* 90.2 (2019), pp. 245–256 (cit. on p. 20).

[DMS04] R. Dingledine, N. Mathewson, P. Syverson. *Tor: The second-generation onion router*. Tech. rep. Naval Research Lab Washington DC, 2004 (cit. on p. 45).

[DPSV04] A. Di Franco, A. Petro, E. Shear, V. Vladimirov. "Small vote manipulations can swing elections". In: *Communications of the ACM* 47.10 (2004), pp. 43–45 (cit. on p. 19).

[Eld12] M. Elder. *Hacked emails allege Russian youth group Nashi paying bloggers*. Feb. 2012. URL: https://www.theguardian.com/world/2012/feb/07/hacked-emails-nashi-putin-bloggers (cit. on p. 16).

[FC11] N. Fielding, I. Cobain. *Revealed: US spy operation that manipulates social media*. 2011. URL: https://www.theguardian.com/technology/2011/mar/17/us-spy-operation-social-networks (cit. on p. 18).

[Fei19a] L. Feiner. *California's new privacy law could cost companies a total of 55 billion to get in compliance*. Oct. 2019. URL: https://www.cnbc.com/2019/10/05/california-consumer-privacy-act-ccpa-could-cost-companies-55-billion.html (cit. on p. 60).

[Fei19b] L. Feiner. *Twitter bans political ads after Facebook refused to do so*. Oct. 2019. URL: https://www.cnbc.com/2019/10/30/twitter-bans-political-ads-after-facebook-refused-to-do-so.html (cit. on p. 55).

[Fer17] E. Ferrara. "Disinformation and social bot operations in the run up to the 2017 French presidential election". In: *arXiv preprint arXiv:1707.00086* (2017) (cit. on pp. 34, 35).

[FH98] P. Ferguson, G. Huston. *What is a VPN?* 1998 (cit. on p. 43).

[FVD+16] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104 (cit. on p. 53).

[GBHF20] A. Goldman, J. E. Barnes, M. Haberman, N. Fandos. *Russia Backs Trump's Re-election, and He Fears Democrats Will Exploit Its Support*. Feb. 2020. URL: https://www.nytimes.com/2020/02/20/us/politics/russian-interference-trump-democrats.html?referringSource=articleShare (cit. on p. 30).

[GD20] V. Gadde, M. Derella. *An update on our continuity strategy during COVID-19*. Mar. 2020. URL: https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html (cit. on p. 41).

[Gel86] L. H. Gelb. *Administration Is Accused of Deceiving Press on Libya*. Oct. 1986. URL: https://www.nytimes.com/1986/10/03/world/administration-is-accused-of-deceiving-press-on-libya.html (cit. on pp. 15, 16).

[GF15] G. Greenwald, A. Fishman. *Controversial GCHQ unit engaged in domestic law enforcement, online propaganda, psychology research*. 2015. URL: https://theintercept.com/2015/06/22/controversial-gchq-unit-domestic-law-enforcement-propaganda/ (cit. on pp. 17, 18).

[Gho18] D. Ghosh. *What You Need to Know About California's New Data Privacy Law*. 2018. URL: https://hbr.org/2018/07/what-you-need-to-know-about-californias-new-data-privacy-law (cit. on p. 60).

[GLB18] A. Gómez-Boix, P. Laperdrix, B. Baudry. "Hiding in the crowd: an analysis of the effectiveness of browser fingerprinting at large scale". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 309–318 (cit. on p. 49).

[Gle20] J. Glenza. *Coronavirus: US says Russia behind disinformation campaign*. Feb. 2020. URL: https://www.theguardian.com/world/2020/feb/22/coronavirus-russia-disinformation-campaign-us-officials (cit. on p. 41).

[Gre17] A. Greenberg. *Hackers Hit Macron With Huge Email Leak Ahead of French Election*. May 2017. URL: https://www.wired.com/2017/05/macron-email-hack-french-election/ (cit. on p. 34).

[Hal18] M. de Haldevang. *Russia's troll farm doubled its budget in early 2018*. Oct. 2018. URL: https://qz.com/1430642/russian-troll-farm-internet-research-agency-doubled-its-budget-in-early-2018/ (cit. on p. 30).

[Hal19] M. de Haldevang. *U.S. Cyber Command operation disrupted Internet access of Russian troll factory on day of 2018 midterms*. Feb. 2019. URL: https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html (cit. on p. 30).

[HGL+19] P. N. Howard, B. Ganesh, D. Liotsiou, J. Kelly, C. François. "The IRA, social media and political polarization in the United States, 2012-2018". In: (2019) (cit. on pp. 21, 23–28, 30).

[HK16] P. N. Howard, B. Kollanyi. "Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU referendum". In: *Available at SSRN 2798311* (2016) (cit. on pp. 35, 36).

[HP18] A. Hern, D. Pegg. *Facebook fined for data breaches in Cambridge Analytica scandal*. July 2018. URL: https://www.theguardian.com/technology/2018/jul/11/facebook-fined-for-data-breaches-in-cambridge-analytica-scandal (cit. on p. 63).

[Ill17] S. Illing. *Cambridge Analytica, the shady data firm that might be a key Trump-Russia link, explained*. Oct. 2017. URL: https://www.vox.com/policy-and-politics/2017/10/16/15657512/cambridge-analytica-facebook-alexander-nix-christopher-wylie (cit. on p. 37).

[JWJ+13] A. Johnson, C. Wacek, R. Jansen, M. Sherr, P. Syverson. "Users get routed: Traffic correlation on Tor by realistic adversaries". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 2013, pp. 337–348 (cit. on p. 50).

[Kas17] D. Kassel. *Mit journalistischem Handwerk gegen Fake News*. Apr. 2017. URL: https://www.deutschlandfunkkultur.de/ard-projekt-faktenfinder-mit-journalistischem-handwerk.1008.de.html?dram:article_id=382926 (cit. on p. 64).

[KLA17] E. Kochkina, M. Liakata, I. Augenstein. "Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm". In: *arXiv preprint arXiv:1704.07221* (2017) (cit. on p. 58).

[Kna20] C. Knaus. *Bots and trolls spread false arson claims in Australian fires 'disinformation campaign'*. Jan. 2020. URL: https://www.theguardian.com/australia-news/2020/jan/08/twitter-bots-trolls-australian-bushfires-social-media-disinformation-campaign-false-claims (cit. on p. 40).

[KSG13] M. Kosinski, D. Stillwell, T. Graepel. "Private traits and attributes are predictable from digital records of human behavior". In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805 (cit. on p. 37).

[LA94] A. Luotonen, K. Altis. "World-wide web proxies". In: *Computer Networks and ISDN systems* 27.2 (1994), pp. 147–154 (cit. on p. 44).

[Lap18] I. Lapowsky. *Cambridge Analytica Could Have Also Accessed Private Facebook Messages*. Apr. 2018. URL: https://www.wired.com/story/cambridge-analytica-private-facebook-messages (cit. on p. 38).

[LCFH18] C. Llewellyn, L. Cram, A. Favero, R. L. Hill. "Russian troll hunting in a brexit Twitter archive". In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. 2018, pp. 361–362 (cit. on p. 36).

[Lee18] C. Lees. "The 'Alternative for Germany': The rise of right-wing populism at the heart of Europe". In: *Politics* 38.3 (2018), pp. 295–310 (cit. on p. 63).

[Mar17] A. Marantz. *The Far-Right American Nationalist Who Tweeted MacronLeaks*. May 2017. URL: https://www.newyorker.com/news/news-desk/the-far-right-american-nationalist-who-tweeted-macronleaks (cit. on p. 34).

[MB18] K. J. Mathews, C. M. Bowman. *The California Consumer Privacy Act of 2018*. 2018. URL: https://privacylaw.proskauer.com/2018/07/articles/data-privacy-laws/the-california-consumer-privacy-act-of-2018/ (cit. on p. 60).

[McC87] J. A. McCredie. "The April 14, 1986 bombing of Libya: act of self-defense or reprisal". In: *Case W. Res. J. Int'l L.* 19 (1987), p. 215 (cit. on p. 15).

[MD05] S. J. Murdoch, G. Danezis. "Low-cost traffic analysis of Tor". In: *2005 IEEE Symposium on Security and Privacy (S&P'05)*. IEEE. 2005, pp. 183–195 (cit. on p. 50).

[Moh17] M. Mohan. *Macron Leaks: the anatomy of a hack*. May 2017. URL: https://www.bbc.com/news/blogs-trending-39845105 (cit. on pp. 33, 34).

[Mue19] R. S. Mueller. *The Mueller Report: Report on the Investigation into Russian Interference in the 2016 Presidential Election*. 2019. URL: https://cdn.cnn.com/cnn/2019/images/04/18/mueller-report-searchable.pdf (cit. on p. 27).

[NN16] G. Neff, P. Nagy. "Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay". In: *International Journal of Communication* 10 (2016), p. 17 (cit. on p. 53).

[NSA12] National Security Agency (NSA). *Tor Stinks*. Document leaked by Edward Snowden in October 2013. June 2012. URL: https://edwardsnowden.com/docs/doc/tor-stinks-presentation.pdf (cit. on p. 48).

[OA20] B. Ortutay, M. Anderson. *Facebook again refuses to ban political ads, even false ones*. Jan. 2020. URL: https://apnews.com/90e5e81f501346f8779cb2f8b8880d9c (cit. on p. 55).

[OB17] D. O'Sullivan, D. Byers. *Exclusive: Fake black activist accounts linked to Russian government*. Sept. 2017. URL: https://money.cnn.com/2017/09/28/media/blacktivist-russia-facebook-twitter/index.html (cit. on p. 21).

[OTo16] M. O'Toole. *In Final Report, Benghazi Committee Finds No New Evidence of Clinton Wrongdoing*. June 2016. URL: https://foreignpolicy.com/2016/06/28/in-final-report-benghazi-committee-finds-no-new-evidence-of-clinton-wrongdoing/ (cit. on p. 26).

[Pay15] W. Payson-Denney. *So, who really won? What the Bush v. Gore studies showed*. 2015. URL: https://edition.cnn.com/2015/10/31/politics/bush-gore-2000-election-results-studies/index.html (cit. on p. 18).

[Per19] B. Perrigo. *British Government Delays Report on Russian Interference in Brexit Vote Until After Election*. Nov. 2019. URL: https://time.com/5717670/uk-russian-interference-brexit-delayed/ (cit. on p. 36).

[PKL03] A. Polyanskaya, A. Krivov, I. Lomako. *The Virtual Eye of Big Brother*. English translation available at http://lrtranslations.blogspot.com/2007/02/commissars-of-internet.html. Apr. 2003. URL: http://www.vestnik.com/issues/2003/0430/win/polyanskaya_krivov_lomko.htm (cit. on p. 16).

[RCV+17] C. Relton, K. Cooper, P. Viksveen, P. Fibert, K. Thomas. "Prevalence of homeopathy use by the general population worldwide: a systematic review". In: *Homeopathy* 106.02 (2017), pp. 69–78 (cit. on p. 40).

[RL17] R. R. Ruiz, M. Landler. *Robert Mueller, Former F.B.I. Director, Is Named Special Counsel for Russia Investigation*. May 2017. URL: https://www.nytimes.com/2017/05/17/us/politics/robert-mueller-special-counsel-russia-investigation.html (cit. on p. 24).

[Rom01] H. Romerstein. "Disinformation as a KGB Weapon in the Cold War". In: *Journal of Intelligence History* 1.1 (2001), pp. 54–67 (cit. on p. 13).

[RSG98] M. G. Reed, P. F. Syverson, D. M. Goldschlag. "Anonymous connections and onion routing". In: *IEEE Journal on Selected areas in Communications* 16.4 (1998), pp. 482–494 (cit. on p. 46).

[San16] D. E. Sanger. *Obama Strikes Back at Russia for Election Hacking*. Dec. 2016. URL: https://www.nytimes.com/2016/12/29/us/politics/russia-election-hacking-sanctions.html (cit. on p. 25).

[SBFF18] E. Sy, C. Burkert, H. Federrath, M. Fischer. "Tracking users across the web via tls session resumption". In: *Proceedings of the 34th Annual Computer Security Applications Conference*. 2018, pp. 289–299 (cit. on p. 48).

[Sch18a] K. Schmehl. *Ein Netzwerk von rund 70 Bots macht auf Twitter Stimmung für Homöopathie und wow, die Recherche war ein wilder Ritt*. May 2018. URL: https://www.buzzfeed.com/de/karstenschmehl/twitter-bots-netzwerk-homoeopathie-fake-user-dzvhae (cit. on p. 40).

[Sch18b] M. Schroepfer. *An Update on Our Plans to Restrict Data Access on Facebook*. Apr. 2018. URL: https://about.fb.com/news/2018/04/restricting-data-access/ (cit. on p. 38).

[Sch19] School of Law, Macquarie University. *MLS Alumni Noelle Martin awarded 2019 WA Young Australian of the Year*. 2019. URL: https://www.mq.edu.au/about/about-the-university/faculties-and-departments/faculty-of-arts/departments-and-centres/macquarie-law-school/news-and-events/departmental-news/news-items/mls-alumni-noelle-martin-awarded-2019-wa-young-australian-of-the-year (cit. on p. 59).

[Sco33] K. Scott. "The political propaganda of 44-30 BC". In: *Memoirs of the American Academy in Rome* 11 (1933), pp. 7–49 (cit. on p. 13).

[Sed14] M. Seddon. *Documents show how Russia's troll army hit America*. 2014. URL: https://www.buzzfeednews.com/article/maxseddon/documents-show-how-russias-troll-army-hit-america (cit. on p. 17).

[Sin14] D. Sindelar. *The Kremlin's troll army*. 2014. URL: https://www.theatlantic.com/international/archive/2014/08/the-kremlins-troll-army/375932/ (cit. on p. 16).

[Smi12] K. Smith. *HOMEOPATHY IS UNSCIENTIFIC AND UNETHICAL: Homeopathy Is Unscientific and Unethical*. Apr. 2012. DOI: 10.1111/j.1467-8519.2011.01956.x. URL: https://doi.org/10.1111/j.1467-8519.2011.01956.x (cit. on p. 40).

[Sör13] O. Sörensen. "Zombie-cookies: Case studies and mitigation". In: *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*. IEEE. 2013, pp. 321–326 (cit. on p. 48).

[SP16] Y. Shveda, J. H. Park. "Ukraine's revolution of dignity: The dynamics of Euromaidan". In: *Journal of Eurasian Studies* 7.1 (2016), pp. 85–91 (cit. on p. 16).

[SSL00] M. L. J. S. C. to Study Bomb Threats in Maine Schools, C. J. Spruce, D. S. Lynch. *Final Report of the Joint Study Committee to Study Bomb Threats in Maine Schools*. 2000. URL: https://digitalmaine.com/cgi/viewcontent.cgi?article=1078&context=opla_docs (cit. on p. 50).

[Sta19] State of California. *Assembly Bill No. 730 - An act to amend, repeal, and add Section 35 of the Code of Civil Procedure, and to amend, add, and repeal Section 20010 of the Elections Code, relating to elections.* 2019. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 (cit. on p. 59).

[Sta85] S. U. C. for State Security (KGB). *KGB, Information Nr. 2955 [to Bulgarian State Security]*. Russian. Committee for Disclosing the Documents and Announcing the Affiliation of Bulgarian Citizens to the State Security and the Intelligence Services of the Bulgarian National Army (CDDAABCSSISBNA-R), f. 9, op. 4, a.e. 663, pp. 208-9. Obtained by Christopher Nehring and translated by Douglas Selvage. Sept. 7, 1985. URL: https://digitalarchive.wilsoncenter.org/document/208946 (cit. on p. 14).

[Str08] D. Stromberg. "How the Electoral College influences campaigns and policy: the probability of being Florida". In: *American Economic Review* 98.3 (2008), pp. 769–807 (cit. on p. 19).

[Stu19] C. Stupp. *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case s*. 2019. URL: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402?mod=hp_lead_pos10 (cit. on p. 52).

[Stu20] J. Stubbs. *Facebook says it dismantles Russian intelligence operation targeting Ukraine*. Feb. 2020. URL: https://www.reuters.com/article/us-russia-facebook/facebook-says-it-dismantles-russian-intelligence-operation-targeting-ukraine-idUSKBN2061NC (cit. on p. 56).

[SWT01] D. X. Song, D. A. Wagner, X. Tian. "Timing analysis of keystrokes and timing attacks on ssh." In: *USENIX Security Symposium*. Vol. 2001. 2001 (cit. on p. 51).

[Thi16] S. Thielmann. *DNC email leak: Russian hackers Cozy Bear and Fancy Bear behind breach*. July 2016. URL: https://www.theguardian.com/technology/2016/jul/26/dnc-email-leak-russian-hack-guccifer-2 (cit. on p. 19).

[TL19] H. Tworek, P. Leerssen. "An Analysis of Germany's NetzDG Law". In: *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression Series, April* 15 (2019) (cit. on p. 64).

[Tob18] M. Tobias. *Comparing Facebook data use by Obama, Cambridge Analytica*. 2018. URL: https://www.wired.com/story/cambridge-analytica-private-facebook-messages (cit. on p. 39).

[Twi18] Twitter Inc. *Update on Twitter's review of the 2016 US election*. 2018. URL: https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html (cit. on pp. 27, 54).

[Uni] United States House Permanent Select Committee on Intelligence (HPSCI). *Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements*. URL: https://intelligence.house.gov/social-media-content/ (cit. on p. 21).

[Uni19] United States Senate Select Committee on Intelligence. *Report on Russian Active Measure Campaigns and Interference in the 2016 U.S. Election Volume 2: Russia's Use of Social Media with Additional Views*. 2019. URL: https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf (cit. on pp. 19, 20, 23).

[Uni87] United States Department of State. *Soviet influence activities: a report on active measures and propaganda, 1986-87*. Includes bibliographies. [Washington, D.C.]: U.S. Dept. of State : [Supt. of Docs., U.S. G.P.O., distributor, 1987, pp. 34–35, 39, 42. URL: http://hdl.handle.net/2027/uc1.31210024732008 (cit. on p. 14).

[US13] S. Usherwood, N. Startin. "Euroscepticism as a persistent phenomenon". In: *JCMS: Journal of Common Market Studies* 51.1 (2013), pp. 1–16 (cit. on p. 35).

[Ush16] S. Usherwood. "Did Ukip win the referendum?" In: *Political Insight* 7.2 (2016), pp. 27–29 (cit. on p. 35).

[Veg10] T. Vega. *New Web Code Draws Concern Over Privacy Risks*. Oct. 2010. URL: https://www.nytimes.com/2010/10/11/business/media/11privacy.html (cit. on p. 48).

[Ver20] L. Verdoliva. "Media forensics and deepfakes: an overview". In: *arXiv preprint arXiv:2001.06564* (2020) (cit. on p. 51).

[VP15] K. P. Vogel, T. Parti. *Cruz partners with donor's 'psychographic' firm*. 2015. URL: https://www.politico.com/story/2015/07/ted-cruz-donor-for-data-119813 (cit. on p. 37).

[Wag00] C. Wagemann. *Das Bild der SPD im"Bayernkurier": die Berichterstattung seit dem Fall der Mauer*. Deutscher Universitäts-Verlag, 2000 (cit. on p. 15).

[WD17] C. Wardle, H. Derakhshan. "Information disorder: Toward an interdisciplinary framework for research and policy making". In: *Council of Europe report* 27 (2017) (cit. on p. 57).

[Web19] M. Webster. *Disinformation | Definition of Disinformation by Merriam-Webster*. 2019 (cit. on p. 13).

[Whi20] Z. Whittaker. *California's new privacy law is off to a rocky start*. Feb. 2020. URL: https://techcrunch.com/2020/02/08/ccpa-privacy-law-rocky-start/ (cit. on p. 62).

[WL12] P. Winter, S. Lindskog. "How china is blocking tor". In: *arXiv preprint arXiv:1204.0447* (2012) (cit. on p. 46).

[Wri16] O. Wright. *EU renegotiations: David Cameron gets 'unanimous agreement' as Britain is given 'special status' in Union*. Feb. 2016. URL: https://www.independent.co.uk/news/uk/politics/eu-renogotiations-david-cameron-unanimous-agreement-britain-europe-a6885206.html (cit. on p. 35).

[YLL19] X. Yang, Y. Li, S. Lyu. "Exposing deep fakes using inconsistent head poses". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 8261–8265 (cit. on p. 52).

[ZRM+18] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al. "A structured response to misinformation: Defining and annotating credibility indicators in news articles". In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 603–612 (cit. on p. 58).

All links were last followed on June 24, 2020.

**Declaration**

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature