



Data Analytics for Social Media Monitoring

**Guidance on Social Media Monitoring and Analysis
Techniques, Tools and Methodologies**

May 2020

Table of Contents

AUTHORS	3
ABOUT NDI	3
ACKNOWLEDGEMENTS	4
INTRODUCTION	5
BACKGROUND	6
WORKING WITH DATA COLLECTION	8
DATA AND NETWORK ANALYSIS	13
IDENTIFYING INFLUENCERS, GROUPS AND ACCOUNTS	20
ANALYZING ACCOUNTS AND CONTENT	25
CONCLUSION	31
APPENDIX I: EXAMPLE API CODE - COLLECTING DATA FROM TWITTER'S SEARCH AND STREAM APIS WITH THE RTWEET PACKAGE	32
APPENDIX II: OSINT TOOLS	34
REFERENCES	35

Authors

Nick Monaco is the Director of the Digital Intelligence Lab (DigIntel) at Institute for the Future. He is an expert in online disinformation and the usage of political bots, specializing in Chinese disinformation. In the course of his work, he has consulted with policy-makers and technologists in both government and industry on how best to combat disinformation and maintain election integrity in countries around the world. He previously worked on these issues at Graphika, a social network analysis and threat intelligence company, and Google's digital rights think-tank, Jigsaw. He is also a research affiliate at the Oxford Internet Institute's Computational Propaganda Project (ComProp).

Daniel Arnaudo is an Advisor at NDI for Information Strategies, covering the intersection of democracy and technology with a special responsibility to develop programs tracking disinformation and promoting information integrity worldwide. Concurrently, he is a Cybersecurity Fellow at the University of Washington's Jackson School of International Studies where he has worked on projects in Brazil, Myanmar, and the United States. Recently, he also collaborated with the Oxford Internet Institute's Computational Propaganda Project. His research focuses on online political campaigns, digital rights, cybersecurity, and information and communication technologies for development.

About NDI

The National Democratic Institute (NDI) is a nonprofit, nonpartisan, nongovernmental organization that responds to the aspirations of people around the world to live in democratic societies that recognize and promote basic human rights.

Since opening its doors in 1983, NDI and its local partners have worked to support and strengthen democratic institutions and practices by strengthening political parties, civic organizations and parliaments, safeguarding elections, and promoting citizen participation, openness and accountability in government.

With staff members and volunteer political practitioners from more than 100 nations, NDI brings together individuals and groups to share ideas, knowledge, experiences and expertise. Partners receive broad exposure to best practices in international democratic development that can be adapted to the needs of their own countries. NDI's multinational approach reinforces the fact that while there is no single democratic model, certain core principles are shared by all democracies.

The Institute's work upholds the principles enshrined in the Universal Declaration of Human Rights. It also promotes the development of institutionalized channels of communication among citizens, political institutions and elected officials, and strengthens their ability to improve the quality of life for all citizens. For more information about NDI, please visit www.ndi.org.

Copyright

© National Democratic Institute (NDI)

Website: www.ndi.org

Copyright © National Democratic Institute for International Affairs (NDI) 2020. All rights reserved. Portions of this work may be reproduced and/or translated for non-commercial purposes with the prior written permission of NDI provided NDI is acknowledged as the source of the material and is sent copies of any translation. Send publication permission requests to legal@ndi.org.

Acknowledgements

The Institute gratefully acknowledges the National Endowment for Democracy (NED) for supporting the creation of this guide. The NED is a private, nonprofit foundation dedicated to the growth and strengthening of democratic institutions around the world. Each year, NED makes more than 1,600 grants to support the projects of non-governmental groups abroad who are working for democratic goals in more than 90 countries. Since its founding in 1983, the Endowment has remained on the leading edge of democratic struggles everywhere, while evolving into a multifaceted institution that is a hub of activity, resources and intellectual exchange for activists, practitioners and scholars of democracy the world over.

It also acknowledges the Institute for the Future (ITF) for their cooperation and partnership in distributing this guide. ITF is a non-profit devoted to civic futures. ITF is the world's leading futures organization. For over 50 years, businesses, governments, and social impact organizations have depended upon ITF global forecasts, custom research, and foresight training to navigate complex change and develop world-ready strategies. ITF methodologies and toolsets yield coherent views of transformative possibilities across all sectors that together support a more sustainable future.

Design and Printing: Ironmark, 2020

Introduction

Social media has become an increasingly important part of the conversations that citizens, candidates, parties and related organizations engage in for political events, elections, referendums as well as votes on bills, strikes and other forms of political activity. It is critical for researchers, election observers, civil society organizations and ordinary people to equip themselves with tools, methods and practices to help in the collection and analysis of data from the online space. Members of international civil society organizations including researchers, program managers, activists and others are engaged in various programs internationally, including as parts of election observation missions, to support local groups to develop their own monitoring capacity, or monitoring on hate speech, political trends and a host of other topics.

This is a guide to help researchers, election observers, technologists and others understand the best practices, tools and methodologies for developing online observation and monitoring for social media networks. It presents an introduction to the relevant concepts to understand when studying these issues, as well as a review of how to build a complete picture of the socio-technical context in a country or region, including the local parties' online presence, social media and internet penetration rates, local media, ethnic and religious divisions and a host of other factors that manifest in the online space.

The contents of this resource feature information about the following key topics:

- **Collaboration:** Researchers must consider potential partners and ways of choosing them, whether in terms of various types of local organizations, international NGOs with expertise in the area, or private firms ranging from smaller examples with expertise in the field to large multinational corporations that control social media platforms and other technologies that reach global networks. A section of this guide will review potential options and considerations for these collaborations, citing examples and noting the benefits and risks of working with different groups.
- **Methodology:** Turning to methodology, the guide examines different methods of data collection, including considerations for different platforms, methods for working with the Application Programming Interfaces (APIs) of each, and for scraping content in different ways.
- **Mapping and Visualization:** This section features guidance on developing and reading maps of networks. The guide introduces key technical terms as well as methods for building maps of the online space; examples of the research from the field; and potential limitations of the maps themselves.
- **Analysis:** This section covers analysis of the different kinds of entities and individuals that shape the conversation. Guidance on this topic includes an overview of individual accounts, influencers and groups, and their roles in the online ecosystem, and also examines ways that news organizations and other outside resources become important sources of content.
- **Content:** Addressing online content itself, the guide looks at different aspects of posts, tweets, and other forms of social media. This section describes how to detect different kinds of networked computational propaganda, ranging from botnets and troll farms to other potential forms of manipulation. It considers various malicious forms of manipulative content ranging from disinformation to hate speech and potential targets.
- **Tools:** To support these research techniques, the guide catalogues and examines the different kinds of tools that are useful in the development of analysis for various aspects of social media monitoring. This inventory includes tools for collection on various platforms, as well as resources for network analysis, data visualization and open-source intelligence research.
- **Responses:** Finally, the guide features a review of potential responses that can be informed and enhanced by these analyses. This includes gathering data for research; developing documentation; and creating reporting mechanisms with platforms, government regulators and election monitoring organizations. The guide concludes with recommendations for building research in future areas and developing critical evaluations of evolving areas of the field.

Background

When entering a new environment, analysts must consider many socio-political actors, networks and groups, as well as technical systems. They must examine various aspects of the informational, social and political environments they are working in, considering the networks and organization of the country itself, the broader region, and its place in the global system. Information is not always passed through online or traditional media networks; much of it exists in word of mouth, through rumors, traditional media, and other methods. However, these discussions now often do permeate online, where they can be better tracked and understood.

Various groups internationally employ tactics to manipulate public perception about candidates and issues, weaken confidence in democratic processes, and confuse voters about polling locations, their registration or the electoral system itself. One of the central purposes of this kind of research is to help illuminate how networks of computational propaganda¹ are operating online, helping to expose how they work, document cases for research, and potentially alert authorities or social media companies to manipulation and abuse online. Detecting automation, false accounts, fake content, bad sources of information and other manipulation could all be objectives of this kind of project to describe computational propaganda in different cases.

Social media analysts looking at information issues in contexts throughout the world need to take many factors into account when they are developing reporting. They should deploy tools, research the laws and institutions governing the information space, and employ offline methods—such as interviews with government officials, parties, media, and candidates—to draw the most accurate picture of disinformation in the election environment.

Disinformation is a difficult subject to capture because it is generally not a very well-defined term. A key component of disinformation is the concept of intent, in that disinformation is passed with the intent to deceive, while misinformation is incorrect content without the necessary intent to falsify. For more on definitions and further background on these subjects, see the Data and Society report *Lexicon of Lies and First Draft's Information Disorder*, as well as NDI's guide on *Supporting Information Integrity and Civil Political Discourse*², which has been translated in Albanian, Arabic, English, French, Russian, Serbian and Spanish. Further sources are noted in the references section. In addition to disinformation, a researcher must consider many different kinds of content, some innocuous, as well as malicious or positive.

It is also important to consider the role of the media, and how they are operating within and influencing the online and social media systems. Look at the major sources of print, radio and television journals and evaluate their market share, links to political parties, involvement of the public and private sector, and how they are influencing the online space. Many have considerable online presence, and particularly when linked with a major political party, can have a huge influence on the editorial line and political preferences of an organization, as well as the propensity to gain viral support, either organic or artificially generated. Polarization in this context can be seen as critical in relation to networks, it can become readily apparent when larger, distinct groups are formed in opposition and coalition, which can also signal the levels of computational propaganda and disinformation present.

Hate speech often includes disinformation to smear the targets of campaigns with false attacks. This particularly targets women, minorities and other vulnerable groups. NDI's Gender, Women and Democracy team's research on *Violence Against Women in Politics (VAW-P)* notes that "When attacks against politically-active women are channeled online, the expansive reach of social media platforms magnifies the effects of psychological abuse by making those effects seem anonymous, borderless, and sustained, undermining women's sense of personal security in ways not experienced by men. Many of the state and non-state actors who perpetrate online VAW-P are mobilizing across transnational networks. The misuse—by states, organizations and individuals—of the very freedoms that the information space is supposed to enable, has become one of the greatest threats to its integrity." (Zeiter et al., 2019, 4) As a result, researchers should be aware of how women are especially vulnerable to these attacks online, and consider ways of documenting and reporting this kind of abuse to platforms. NDI has piloted the use of lexicons of

hate speech terms to study social networks with case studies in Colombia, Indonesia, and Kenya. The methods for the development of these lexicons and the case studies which is detailed in the report on this research “Tweets that Chill: Analyzing Violence Against Women in Politics.”

When developing strategies for data collection, consider how these information operations using hate speech and computational propaganda are operating within the context you are working in. The subsequent contents of this guide will help analysts learn how to identify these online campaigns, networks and users, but researchers should understand broadly what kind of computational propaganda, harmful speech or other patterns they are looking for before proceeding.

It’s also important to have a broad understanding of both the local regulatory environment (including electoral and campaign finance regulation, which can help inform reporting), and the rules of content moderation and policies that govern the platforms being studied. This is so researchers can design their studies legally and ethically, and alert companies--and potentially governments--to abuse and other illegal and negative actions online. This also helps researchers to better understand the countries they are studying.

Understanding the terms of service for the various social media companies represents a critical aspect of this reporting. Key standards to understand for Facebook, Twitter and YouTube with links to the policies are listed below.

Facebook Community Standards	<ul style="list-style-type: none"> ▪ Substantive info on reporting different scenarios, here. ▪ “Report an Imposter Page of a Public Figure,” here. ▪ “How to Report Things” broken down by different type of post, here. ▪ “How do I mark a post as false news?,” here.
Twitter Rules	<ul style="list-style-type: none"> ▪ General overview of report violations, here. ▪ How to report a tweet, abusive account, or individual message here. ▪ Reporting an impersonation account, here. ▪ Report spam instructions, here.
YouTube Community Guidelines	<ul style="list-style-type: none"> ▪ How to report inappropriate content, broken down by type of post, here. ▪ “Report a YouTube search prediction,” here. ▪ “Other reporting options,” here. ▪ “How to report spam or deceptive content,” here (end of page). ▪ YouTube reporting tool, here.

Some companies, such as Facebook, require users to identify themselves with real information, so simply identifying an account that is not a real person, or a group connected to that false account, can lead to a takedown. Others, such as Twitter, do not prohibit anonymity but do have prohibitions against hate speech or artificial amplification that can be identified and reported through research. It is worth familiarizing yourself with the various codes and mechanisms for reporting content, which are outlined in the table above.

Reporting is important, as is documenting the campaigns, accounts, and relevant content so that reports can be verified. Consider using systems that can be easily backed up and consulted, in terms of annotation and workflows. As noted, Facebook’s Community Standards offer mechanisms for reporting false accounts and negative forms of content, and to request fact checkers verify and potentially take down harmful content. Twitter’s Rules allow for different forms of reporting for impersonation, spam and certain forms of hateful or otherwise prohibited speech. YouTube’s policies are focused on media and copyright, as well as explicit, hateful or other harmful forms of speech. As the owners of YouTube, Google’s terms of reference give you an idea of how you can not only report accounts but also investigate accounts linked to the video streaming service.

In terms of governmental regulations, analysts should consider data protection laws such as the European Union's General Data Protection Regulation, which contains aspects that cover the collection of personally identifiable information in Europe or about Europeans anywhere, as well as companies that operate there.³ Collection of data from private groups, users and networks could abrogate such laws as well as the terms of service of the platforms.

Networks such as the Design 4 Democracy Coalition (D4D Coalition) can help advocate for democratic principles at technology companies—for instance, to bring larger scale influence campaigns for electioneering, hate speech or other purposes to the attention of the platforms. The D4D Coalition is comprised of international and national civil society organizations engaging with tech companies to integrate and support democratic principles, including in the context of content moderation, policy development, and product considerations. The Coalition (with leadership from NDI, the International Republican Institute, the International Foundation for Electoral Systems, and International IDEA) links civil society and democracy stakeholders in diverse contexts with many of the most influential tech companies (including Facebook, Microsoft and Twitter) to encourage information-sharing and to advance strategies for promoting information integrity and protecting democratic processes. Efforts such as these can be supported and enhanced by solid documentation, analysis, and reporting using the tools, methods and tactics described here.

For more details on developing strategies for data analysis on social media in elections, see NDI's guidance document on Disinformation and Election Integrity⁴, as well as Supporting Democracy's Guide for Civil Society on Monitoring Social Media for Elections.⁵ These guides cover methodologies and regulatory considerations for observers in terms of social media monitoring and the possibility of integrating data collected online into traditional election observation missions.

Working with Data Collection

Collecting data is the first step to rigorous analysis of online activity around an election. As an on-the-ground analyst, the first step to collecting data is a survey of the online/social media landscape in the region you are observing. Important questions to consider include:

- **What platforms are most popular in the region? What are different platform penetration rates?** [Internet World Stats](#), the [International Telecommunications Union](#), Freedom House's [Freedom on the Net report](#), or [Facebook itself](#) can be good resources here. A country with high rates of engagement on Facebook but low penetration rates on Twitter (such as Taiwan), would likely yield the most valuable insights on Facebook.
- **What websites are popular for news? Which of these are legacy media? Which of these are more recent creations?**
- **What hashtags are most relevant to the election in question? Similarly, what official accounts represent parties, candidates and their campaigns in these elections?** Oftentimes a candidate will have more than one account or page relevant to an election - such as a personal Twitter account and an official campaign Twitter account. Making a list of these is a good first step for collecting relevant data.

After discerning the most important social media platforms and websites in the country, you're ready to start collecting data on relevant platforms and websites. In this section, we'll examine the options available to researchers for data collection.

³ <https://gdpr.eu/>

⁴ <https://www.ndi.org/publications/disinformation-and-electoral-integrity-guidance-document-ndi-elections-programs>

⁵ Supporting Democracy is implemented by a consortium composed of SOFRECO, Democracy Reporting International (DRI) and NDI. <https://democracy-reporting.org/wp-content/uploads/2019/10/social-media-DEF.pdf>

Collection methods

When collecting social media data online, researchers have several options available to them. The three main options are using third-party collection tools, directly engaging with a platform's API, or engaging in "web scraping". We'll explore the details and differences between these three methods in this section.

Third-Party Tools (Indirect Access)

For researchers who are on a tight timeline or do not have the capacity to interact with websites through computer code, third-party tools can be a useful option. These tools typically interact with the APIs of one or more target platforms invisibly in the background, and present data in a user-friendly, graphical layout such as a dashboard. For Facebook, which [does not currently allow external researchers or businesses to use its API](#), [CrowdTangle](#) is one of the best options for monitoring reach of pages, groups and URLs on Facebook, and also shows users the reach of URLs on Twitter, Instagram and Reddit. The CrowdTangle Extension⁶ allows a user to see a real-time estimate of the number of reactions a post, page or URL has garnered on these platforms, which can be a useful way of monitoring trending content across multiple platforms on a day-to-day basis.

Certain third-party tools, such as Sysomos and Brandwatch, require expensive paid subscriptions to use, but offer access to a vast amount of data that can be useful for monitoring hashtag campaigns and other trending content on social media.

In addition to being visualized on a web browser, data from these tools can often be exported to a machine-readable file, such as a Comma Separated Values (CSV) file, which can in turn be manipulated by a data scientist to query the data in new and useful ways.

Facebook: Sysomos, Brandwatch

Twitter: Twitonomy

Direct Access - APIs and Web Scraping: What's the difference?

For more direct interaction with platforms and websites, researchers have two available options: using an Application Program Interface (API) or collecting the information directly from the web page's source code, a practice known as "web scraping". The difference between APIs and web scraping is important to note - pulling data from APIs is in most cases legal and ethical since the API data is deliberately regulated by platforms and structured so as not to violate users' rights. Web scraping is, in many cases, a violation of Terms of Service, and is tougher to regulate. In many cases, web scraping can be illegal.

It is important to note the difference between these two methods of data collection. It is also important to know that you will occasionally hear researchers themselves erroneously refer to data that was retrieved from an API as being "scraped". The distinction is important - not only for practical reasons like saving time and effort, but also for the ethical and legal violations that can come from web scraping.

Keeping this basic analogy of the difference between the two approaches in mind, let's dive into the details of APIs and web scraping.

APIs

For researchers interested in a more hands-on approach to data collection, many platforms have a more direct form of access to data in the form of application programming interfaces (APIs). Generally speaking, APIs are a way

⁶ CrowdTangle is currently available to academics and researchers on a selective basis. You and your team can apply at [CrowdTangle.com](#). The full version includes current and historical data on Facebook and Instagram. A free CrowdTangle extension is also available at the website. This extension takes a URL as input and offers the 500 most recent public posts that have garnered any significant traction citing the URL on Facebook, Instagram, Reddit and Twitter. Both the full platform and the CrowdTangle Extension can be useful for investigations.

for users to easily interact with a website or social media platform through computer code. This enables a user to quickly process much more data, and in turn generate deeper insights into online activity, than they would otherwise be able to do manually.

Types of APIs

Two API features in particular are relevant to be aware of before data collection begins: openness and timeframe.

- **Openness:** APIs come in different forms - open APIs allow data to be collected by anyone, while authenticated APIs require a user to undergo some form of verification before permitting data collection.
 - **Open APIs:** Venmo, the app used for electronic payments in the US, has a [public API](#) that allows anyone to view a number of the most recent public transactions on the app. You can find lists of Open APIs all around the internet - such as [this list](#) on Github. [Any-api.com](#) also has a list of several APIs that are publicly available to interested users, many of which are open.
 - **Authenticated APIs:** Most APIs you'll be dealing with for social media data gathering (Twitter, Reddit, etc.) require a user to be authenticated before they can begin gathering data.
- **Timeframe:** Websites and platforms also typically structure their APIs differently depending on time.
 - **Gathering historical data:** Most APIs allow you to collect some form of historical data on their sites - Twitter and Reddit notably allow this. This form of API, let's call it an *historical API*, allows you to retroactively pull data that was generated *before* you made the query. Importantly, data posted after you made the query is not available for collection.
 - **Streaming real-time data:** Ingesting and downloading data as it occurs in real time is a process referred to as *streaming*. If Twitter is a relevant platform in the election or period you're planning to monitor, streaming is likely to be your best option for data collection. When streaming data, you collect Tweets in real time according to a specified query (e.g. all tweets using the hashtag *#election*, or all tweets mentioning accounts of interest, citing URLs of interest, etc.)

API Collection Limitations

In the interest of preserving user privacy and safety, most social media platforms limit the amount of data any one user is allowed to collect. We explore a few of those limitations here to familiarize you with issues you may encounter when collecting data.

- **Volume limitations:** Most APIs limit the volume of data that a given user is able to collect. For example, when streaming real-time data on Twitter, the platform caps the amount of data a user is able to collect to 1% of global streaming data.
 - **Rate limits:** Rate limits are the most common type of volume limitation you will run into when pulling API data. Most APIs have rate limits to ensure that a single user or application cannot download an excessive amount of data (as defined by the particular platform or website). For example, Twitter [limits](#) the number of Tweets a single user can download within any 15-minute window.
- **Removed data limitation:** As touched on briefly above, platforms tend to remove content that violates the specified rules and regulations of use - these regulations have several different names (community standards, terms of service, etc.). This note is particularly important for election contexts in which nefarious behavior is likely to occur - in particular, streaming data on Twitter enables you to gather and preserve data on nefarious actors in real time that may be removed from the platform later. Once removed, data on these actors is unavailable. If your team wants to capture bad actors, disinformation and other content for analysis later, streaming in real time maximizes your chances of doing so.
- **Data type limitations:** Similarly, most platforms will limit the type of data that a user can collect from its platform. Twitter's API will allow you to collect certain information about a target user (such as their number of posts, number of followers, and account creation date), but it will not allow you to access other kinds of

restricted data (such as a user's most frequently used IP address). Facebook does not currently allow for researchers to gather information about other users or pages through its API at all. It's useful to be familiar with what kind of data can be collected on target platforms when designing a research project.

- **Data time limitations:** Twitter limits users to data having occurred in the past 7-9 days when collecting historical data through the Twitter Search API. Data that is older than 7-9 days on Twitter can be purchased from data providers such as GNIP, but cannot be collected through standard API access.

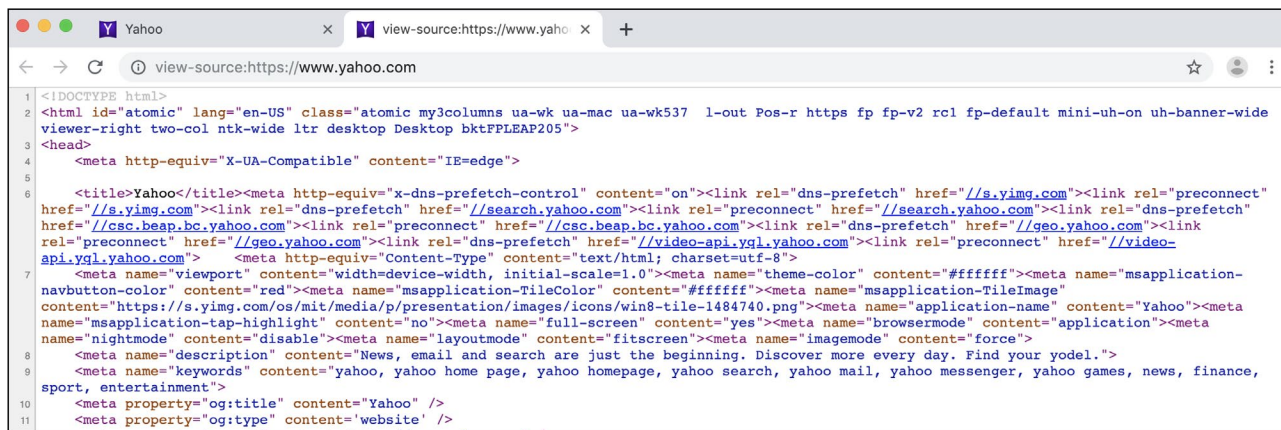
An exciting aspect of APIs is that they are not language-specific. API data can be pulled using Java, Python, R, Ruby, Perl, or any other programming language that you or your tech team may prefer. While this is the case, there are specific packages in popular programming languages that simplify the process of querying the API by handling some of the complexities for you.

Example code with explanations showing how to use R's rtweet package to collect data from Twitter's Search and Streaming API is available at the end of this field guide in the Appendix: *Example API code - Collecting Data from Twitter's Search and Stream APIs with the Rtweet Package*.

Web Scraping

While APIs offer a streamlined way of gathering data from a platform or service, they are not the only option for gathering data. *Web scraping* is the process of extracting the source code from a target website and mining out relevant data. Every page on the Internet is the result of the source code that it is made up of - HTML and other dynamic web languages, scripts, hyperlinks and media sources. The process of a browser taking code text and turning it into a visual and interactive webpage is called *rendering*. You can view the underlying source code for any webpage in most modern browsers - Google Chrome, Mozilla Firefox, Safari, Brave and Opera all have this feature.

For example, when using Google Chrome, you can right-click on any webpage that has been loaded (or "rendered") in your browser and click on the "View Page Source" option. Chrome will open a new tab that shows you the HTML and CSS code that are used to load the webpage you are viewing. An example of this from yahoo.com is shown below.



```
1 <!DOCTYPE html>
2 <html id="atomic" lang="en-US" class="atomic my3columns ua-wk ua-mac ua-wk537 l-out Pos-r https fp fp-v2 rcl fp-default mini-uh-on uh-banner-wide
viewer-right two-col ntk-wide ltr desktop Desktop bktPFLEAP205">
3 <head>
4   <meta http-equiv="X-UA-Compatible" content="IE=edge">
5
6   <title>Yahoo</title><meta http-equiv="x-dns-prefetch-control" content="on"><link rel="dns-prefetch" href="//s.yimg.com"><link rel="preconnect"
href="//s.yimg.com"><link rel="dns-prefetch" href="//search.yahoo.com"><link rel="preconnect" href="//search.yahoo.com"><link rel="dns-prefetch"
href="//csc.beap.bc.yahoo.com"><link rel="preconnect" href="//csc.beap.bc.yahoo.com"><link rel="dns-prefetch" href="//geo.yahoo.com"><link
rel="preconnect" href="//geo.yahoo.com"><link rel="dns-prefetch" href="//video-api.yql.yahoo.com"><link rel="preconnect" href="//video-
api.yql.yahoo.com">
7   <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
8   <meta name="viewport" content="width=device-width, initial-scale=1.0"><meta name="theme-color" content="#ffffff"><meta name="msapplication-
navbutton-color" content="red"><meta name="msapplication-TileColor" content="#ffffff"><meta name="msapplication-TileImage"
content="https://s.yimg.com/os/mit/media/p/presentation/images/icons/win8-tile-1484740.png"><meta name="application-name" content="Yahoo"><meta
name="msapplication-tap-highlight" content="no"><meta name="full-screen" content="yes"><meta name="browsermode" content="application"><meta
name="nightmode" content="disable"><meta name="layoutmode" content="fitscreen"><meta name="imagemode" content="force">
9   <meta name="description" content="News, email and search are just the beginning. Discover more every day. Find your yodel.">
10  <meta name="keywords" content="yahoo, yahoo home page, yahoo homepage, yahoo search, yahoo mail, yahoo messenger, yahoo games, news, finance,
sport, entertainment">
11  <meta property="og:title" content="Yahoo" />
    <meta property="og:type" content="website" />
```

Screenshot of the HTML and CSS source code underlying yahoo.com (snapshot taken in early August 2019). Google Chrome and other modern browsers allow users to view the source code of any website they visit - this source code is what is retrieved when a website is "scraped".

Ethical concerns when web scraping

Web scraping can be a useful tool for analyzing web pages. It should be noted though, that web scraping is a violation of most social media platforms' terms of service, and it can be easy to break the law when scraping web pages. For this reason, it's important to always consider user privacy, relevant terms and conditions and relevant laws when scraping websites. It is always important to make sure your data collection process is both ethical and legal. **Many researchers only gather data through APIs and choose not to scrape websites for these reasons.**

When using data collected by other teams or tools - a frequent scenario for many teams - it is also paramount to make sure the data you are working with was obtained ethically. For example, if the data was procured illegally through scraping or hacking, it is inadvisable to use it for a research project for legal, ethical and political reasons, and this kind of project can lead to repercussions from both governmental and corporate authorities. These considerations are important to take into account before beginning data analysis.

Platform differences: *Summary Table*

	Historical API available	Streaming API available	Third-party data collection tools
Twitter	Yes	Yes	CrowdTangle (extension)
Facebook	No	No	CrowdTangle
Gab	Yes, through Pushshift.io	No	No
Instagram	No	No	CrowdTangle
Reddit	Yes	Yes (certain packages have this functionality - e.g. Reddit SSE)	CrowdTangle (extension) Pushshift.io
YouTube	Yes	No	
Telegram	Yes	No	Telethon Pushshift.io
Vkontakte	Yes	No	
WhatsApp	Yes - the WhatsApp business API enables automated communications from businesses to customers. It is not typically used for the type of analysis we discuss in this guide.	No	Several third-party tools enable statistical analysis or visualization of WhatsApp chats ⁷ . <ul style="list-style-type: none"> ▪ ChatAnalyzer ▪ WhatsApp Chat Analyzer ▪ Chatilyzer ▪ WhatsAppAnalyzer

Useful tools for data collection in an election integrity context:

- **Twitter API packages:**
 - R packages: [rtweet](#), [twitterR](#)
 - Python packages: [python-twitter](#), [tweepy](#)
- **CrowdTangle** - CrowdTangle and the CrowdTangle Extension are currently the best tools for analyzing Facebook and Instagram.
- **Pushshift.io** - a site that archives data from social media platforms. Reddit, Gab, Twitter and Telegram. The founder and operator of Pushshift, Jason Baumgartner, obtains his data through API access, making the use of the data safe from an ethical perspective.
 - **Reddit streaming tool:** https://github.com/pushshift/reddit_sse_stream

⁷ It is important to be clear about the privacy implications of using these tools. In particular, you and your team want to ensure that private chats are not being exposed to or saved by a third-party when you encounter WhatsApp analysis tools.

- **MIT Media Cloud** - MIT Media Cloud is a news aggregation tool that can be useful for exploring coverage of topics of interest in diverse outlets. Describing the tool, the [website](#) reads “We aggregate data from over 50,000 news sources from around the world and in over 20 languages including Spanish, French, Hindi, Chinese, and Japanese. Our tools help analyze, deliver, and visualize information about media conversations on three primary levels: attention and coverage peaks of issues, network analysis, and clustered language use.”

Data and Network Analysis

Building visual representations of social networks and analyzing the relationships within is a process known as social network analysis (SNA). This process is also commonly referred to as building a social network map, or “mapping” a social network. While it is not always necessary to use social network analysis to understand the online media sphere around an election, it can be a useful way of generating informative insights about influence within a given swath of a social media community, and it can be a useful way to visualize that community.

At its core, making a social network map is a process that consists of 5 steps:

1. Data Collection
2. Deciding what Relationship to Map
3. Data Pruning
4. Map Generation
5. Map Analysis

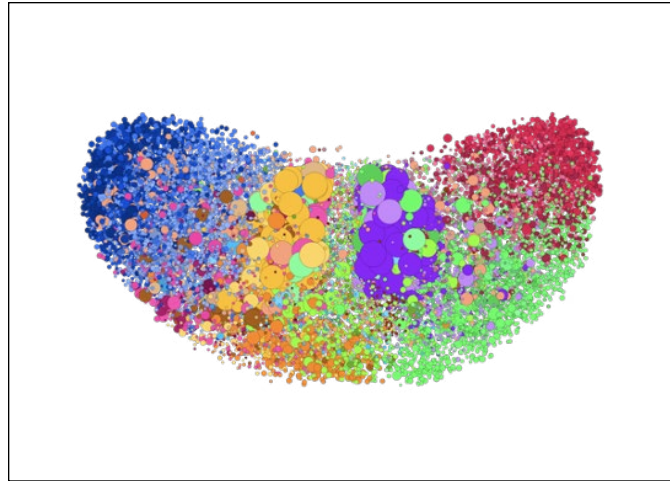
We’ll discuss each of these steps in greater detail below.

Basic Terminology

When discussing and analyzing social network maps, there are a few terms that are useful to know to help understand the typical vocabulary and practices of social media monitoring. Networks are extremely useful because they can represent many relationships in many different contexts. While maps of social networks are probably the most familiar, networks can be used to model the spread of diseases and viruses, map neural activity in the brain, or represent possible routes of traveling from one city to another. While this applicability of networks to so many different domains is useful, it also produces a need for abstract language to talk networks regardless of the application space. This section briefly introduces you to a few basic terms that are useful for discussing networks.

The most relevant ones are *graph*, *node* and *edge*.

- **Graph** - *Graph* is the computer science terminology for a network that is composed of *nodes* and *edges*. It’s important to know this word as you may encounter it in tools that make network maps, or in discussions of those tools. You can consider the word *graph* to be roughly synonymous with *network*.
- **Node (or vertex)** - *Nodes* are the elements that make up a network. One key thing to know about nodes is that what a node represents is different depending on the map/visualization you are looking at. A node in [Lawrence Alexander’s network map of pro-Kremlin websites](#) represents a domain, while each circular node in Graphika’s 2018 map of the American Twitter political landscape represents a Twitter account.



This [Graphika map of the 2018 US Political Spectrum on Twitter](#) illustrates nodes. Each circle in the map is a node above representing an individual Twitter account. Nodes and edges are the two main building blocks of networks.

- **Edge (or arc)** - Edges are the connections between nodes in a network, most commonly represented as a simple line between two nodes. These connections can represent a variety of things. In a model of disease contagion, edges may represent the spread of a virus from one host to another. In a graph of airports in the United States, edges between two nodes (airports) may represent a direct flight available between the two airports. Edges can be directed or undirected⁸, and they can have a numeric value associated with them⁹ (often referred to as *weight*).
- Edges in social media graphs are likely to represent one of a few things: Following relationships, retweets or likes. Most networks you see on Twitter will be *follower networks*¹⁰ - in which the *edge* displays that one user follows another - or *retweet networks*.

Making a Social Network Map: Steps Involved

In this section, we'll define and examine the five steps involved in making a social network map from which to generate insights.

1. **Data Collection** - Examined in the previous section, this step involves collecting data relevant to a local election through a social media API or third-party tool. Once data is collected, you have the base data pool you'll need to make a social network visualization. It is important to realize that *only parts of this data will be used in the generation of the map* - the same base dataset can be used to generate all kinds of maps. How to choose what kind of network/relationships you are most interested in and what pieces of data are most relevant is the work of steps 2 and 3.
2. **Deciding what relationship to map** - In your social network map, you're likely to have each "node" (i.e. circle in the map) represent a Twitter account or Facebook page. However, the relationships between these nodes (liking relationships, following *relationships*, retweets, etc.) are what give the network of relevant nodes its structure. You can think of this relationship as determining the skeleton of the map - from that skeleton, we extract relevant relationships about the nodes in the map. In graph theory - the field of computer science that deals with networks and mapping - these connections are also referred to as *edges*.

⁸ For example, the graph modeling viral contagion discussed above may have a direction associated with the spread of infection, the edge would move *from* the infector to the infected in this case, and would be depicted with an arrow.

⁹ In the airport graph example above, values of edges could be the mile distance between airports. Another possible edge value for this graph would be the time it takes to fly from one airport to another.

¹⁰ For examples of follower networks see a network of accounts promoting anti-vaccine messages in Wired [here](#).

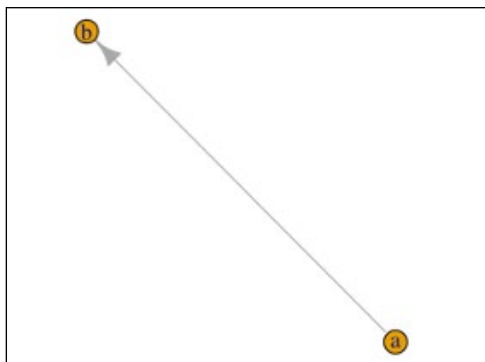
3. **Data Pruning** - If you collect a large amount of data, you will almost always end up with more data than you can map. Affinio, NodeXL, Gephi and other map generation tools tend to operate best in the range of a few thousand nodes. Mapping at a higher scale is generally too computationally expensive to be useful - for this reason, it's necessary to figure out the most relevant network for your questions. The process of cutting out extraneous data is often called *pruning* in computer science - you may also hear this referred to as *network reduction* or *dimensionality reduction*. Some tools (e.g. Graphika) do this network reduction work for you, but you can also make thresholding decisions on your own with tools such as Gephi. Some examples include only mapping nodes who have used an election related hashtag more than once, or only including nodes that have a connection to five or more nodes in the network.¹¹
4. **Map Generation** - After you've decided what relationship to map (edges) and what nodes will be in your network through network reduction, you are ready to make your map. Generally at this stage, you will have to put your relevant data into a format that is readable for the tool you are using (e.g. a CSV file, a [.graphml](#) file, or a [.gexf](#) file for Gephi), and read it into your software. After this, most of the hard work is done for you. There are plenty of free and useful [tutorials](#) on YouTube for generating network maps with Gephi¹² and other open-source tools.
5. **Map Analysis** - Once your map is generated, you can customize the visuals and move on to analysis. Normally, measures of *centrality* are key for understanding influence in a network. [Analyzing Social Networks](#) (Borgatti, Everett and Johnson 2019) has an entire chapter dedicated to different types of centrality that is worth consulting.¹³

Examples of Types of Networks

As mentioned above, there are several kinds of networks that can be generated from a given set of data. The key determining factor for what kind of network you are mapping lies in the *relationship* you have chosen to map. On Twitter, two common types of directed networks are frequently used to analyze data: following networks and retweet networks. Both types of networks can be useful for analyzing data and influence.

Following Networks

In a "following network", nodes are Twitter accounts and the connections between them represent following relationships. Normally these connections are directional (proceed from one node to another in a particular direction). You can think of these as lines that mean "*is following*". For instance, the graph below shows two nodes, A and B, and shows that "A is following B".



Depiction of a one-way following relationship - in this diagram, A is following B. This is depicted by a single directed line (or arrow) from A to B.

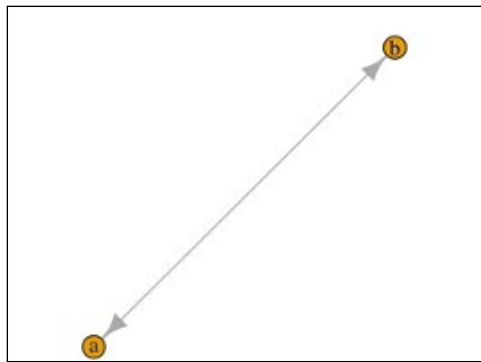
Notice that in this graph, there is only a following relationship in one direction - that is, A is following B, but B is *not* following A. If there were a mutual following relationship, in which A and B were both following each other, a network

¹¹ This kind of pruning is known as *k-core reduction*.

¹² Gephi also offers a [free PDF tutorial](#) and other learning materials on its official site - [gephi.org](#).

¹³ Chapter 10, Centrality. This book is a great text for learning the basics about networks and network analysis methods. As always, there are also all sorts of great free and open-source tutorials online.

depiction would show arrows in both directions, such as in the figure below.



Depiction of a mutual following relationship: A follows B, and B follows A.

In networks generated on Twitter data around an election, you are likely to have many more than two nodes in your relevant network. The examples above serve to give an idea of the basic building blocks out of which a complex network is constructed.

Advantages and Disadvantages of Following Networks

Following relationships on Twitter in particular are somewhat long-term, permanent relationships - users do not often unfollow other users. For this reason, following networks represent a longer-term view of network dynamics and information flows than retweet networks. On the other hand, users tend to have more than one interest on Twitter - this opens up the possibility that some of the users in your follower network may not be as relevant to the content you're interested in as you'd like. Automated networks often have more monothematic focuses, for instance to support a particular party, candidate or issue, but still often vary in content. It is certainly one aspect to consider when analyzing accounts for automation or other forms of coordinated activity.

It would be entirely possible, for instance, in a hypothetical map of the US political spectrum, to have a swath of the network that primarily tweeted about pop culture and music, but that was significantly connected to a network of primarily political accounts. These advantages and disadvantages are useful to keep in mind when deciding whether a follower network is of interest for your task.

Retweet Networks

Retweet networks are another kind of relationship on Twitter - connections between nodes in these networks represent *who retweeted whom* within the collected data. In this regard, these maps are more content-oriented than follower networks. A depiction of edges within this type of graph is shown below.

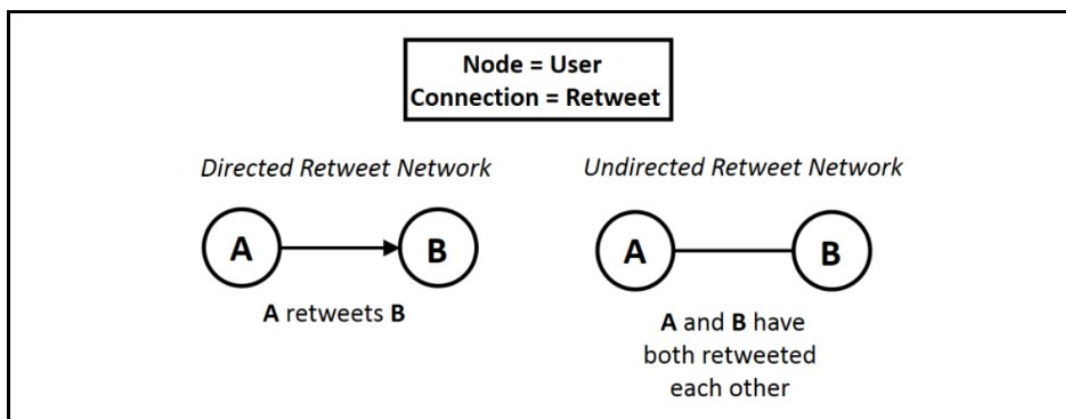
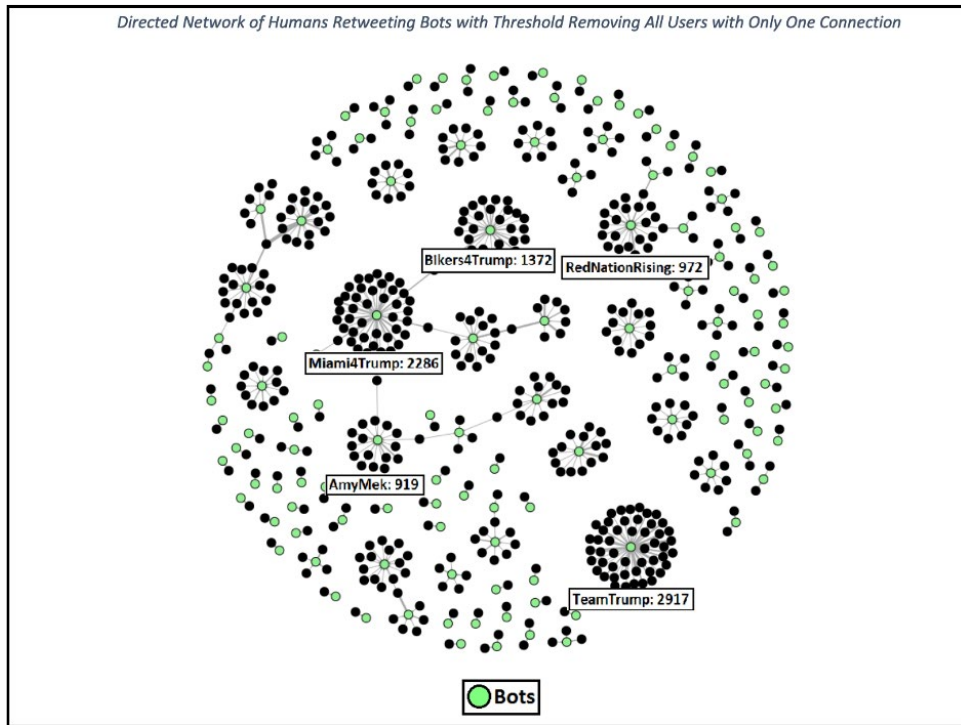


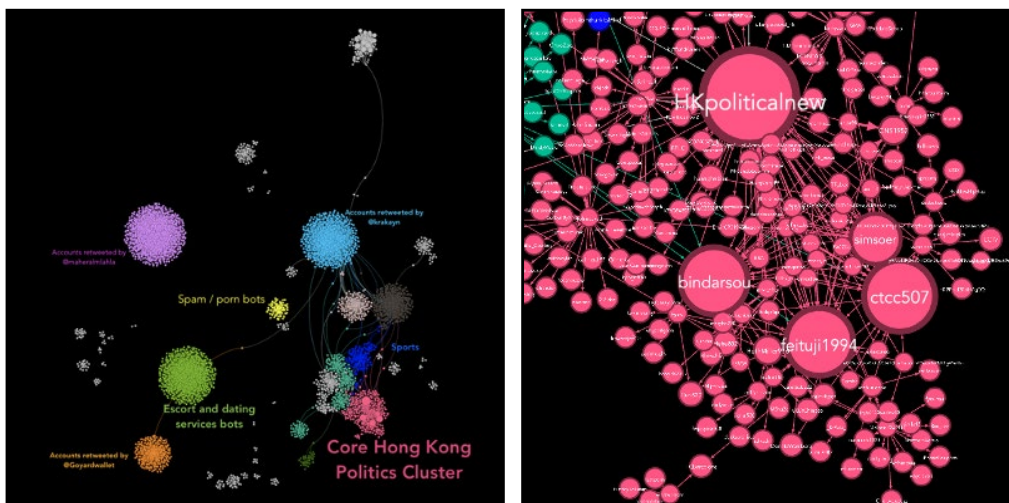
Diagram depicting nodes and edges within a retweet network. (Source: Samuel Woolley and Douglas Guilbeault (2017). Computational Propaganda in the United States: Manufacturing Consensus Online. Available [here](#).)

Advantages and Disadvantages of Retweet Networks

These networks are somewhat more ephemeral than follower networks, as retweet networks represent a snapshot in time - who was retweeting whom within a given timeframe. In this regard, they are an accurate depiction of influence dynamics within a short timeframe - such as a dedicated hashtag campaign or the days preceding an election.



A retweet network generated from Twitter data collected around the 2016 US presidential election. In this network, nodes representing bot accounts are green, and human accounts are black. This retweet network depicts retweets between users and shows that humans significantly retweeted bot content in the 2016 presidential election. (Source: Samuel Woolley and Douglas Guilbeault (2017). Computational Propaganda in the United States: Manufacturing Consensus Online. Available [here.](#))

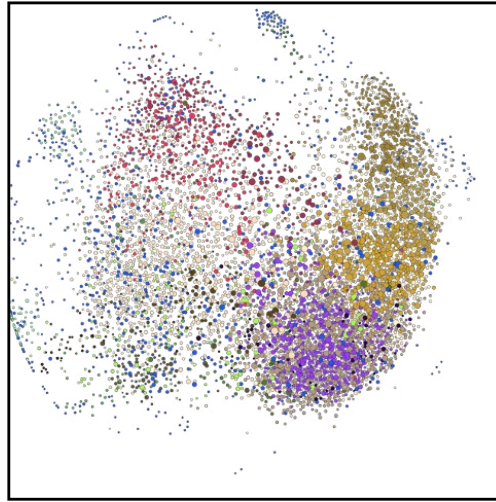


Images of a retweet network of Chinese government accounts spreading disinformation to discredit pro-democracy protests in Hong Kong from June-August 2019. This retweet network was compiled from [Twitter's information operations archive](#)¹⁴. These images represent the entire retweet network (left) and a zoomed-in view of the main political cluster targeting Hong Kong (right). (Source: Digital Intelligence Lab at Institute for the Future report, available [here.](#))

14 Specifically from the August 2019 release of Tweets attributed to the Chinese government.

Mention Networks

Mentions maps are another common kind of map you are likely to encounter. The idea is much the same as retweet maps - edges represent “mentions” - instances in which one user mentions another. Using Gab data assembled by Pushshift.io, Graphika generated a mentions network map to analyze conversations and communities on the social platform Gab - a social network that is very similar to Twitter in its usage and construction.



Graphika's mentions map of Gab users. This map was compiled from Pushshift.io's publicly available Gab dataset, which spans from August 2016 - late October 2018.

Following, retweet, and mention networks are but a few of the options you have for generating networks from Twitter data - there are certainly other possibilities. On other platforms, such as Facebook or Gab, other relationships may be possible. One can imagine a network map of public pages and liking relationships between them, for instance.

Data Collection and Network Mapping Limitations

As with any tool, network mapping has certain limitations, as does the data collection process that precedes it. The two most important limitations to keep in mind are scale and time. As mentioned in the data collection section, all the available tools for generating network maps currently have scale limitations, whether free or paid. This is because as the number of nodes in a network grows, the number of possible connections grows *exponentially*¹⁵. This amount of data is taxing, even for computers, and thus you are unlikely to find a tool that can generate a map of greater than around 15,000 nodes¹⁶. It would be wrong to see this limitation exclusively as a problem, however - an important part of data science and data analysis involves choosing which parts of a large dataset are most likely to yield useful insights. Put another way, analyzing data properly involves choosing which parts of the data are most worth investigating - in this way, data science is as much an art as it is a science. This task, which is a central part of data investigation, analysis and understanding, is not always made easier by having more data.

The second limitation you are likely to encounter with data collection and network mapping is a time limitation, specifically when gathering historical data. The standard Twitter Search API only allows historical data collection going back a week from the time of query. Data that is older than 7-9 days must come from another source - such as a (paid) tool that enables farther-reaching historical collection, or purchasing data from a data provider, such as [GNIP](#). While analysis of historical data can be useful in certain cases, purchasing historical data can get expensive quickly. For this reason, **the best solution is always to collect all relevant data in real time**. If you or a member of your team are interested in streaming or collecting data around an election, it is always best to begin collecting as soon as you have a clear idea of what you'll be interested in.

¹⁵ For those interested in the math, a network of n nodes has $n*(n-1)/2$ possible connections.

¹⁶ Graphika's mapping software is currently able to map at the greatest scale, enabling the visualization of up to 5.5 million nodes

Closed and Encrypted Networks

A challenge facing media monitoring efforts in recent years is the popularity and widespread adoption of encrypted messaging applications such as WhatsApp, Telegram and Signal. Increasingly, the media that citizens rely on to keep abreast of political developments during elections and other major political events is occurring on these platforms. Countries such as Brazil, India and Mexico have seen a marked uptick in political messaging on WhatsApp, for instance. While the usage of encrypted messaging networks is undoubtedly advantageous for the privacy, security and digital rights of citizens, it poses new challenges for understanding the dissemination of political information online.

At present, the best methods for media monitoring efforts of closed networks rely on manual fact checking - teams of experts who monitor relevant WhatsApp channels individually or as a group, fact check stories and distribute the results of these efforts publicly. Such efforts have been successful in several cases such as La Silla Vacía¹⁷ WhatsApp detector in Colombia, as well as work by Verificado¹⁸ in Mexico, and the Center for Democracy and Development in Nigeria.¹⁹

Another noteworthy effort is the [Cofacts bot](#) in Taiwan. Johnson Liang and a team of developers working with the g0v movement devised a way to combine manual fact-checking and automated distribution to help Taiwanese citizens check whether a story is false or not. Users can add the Cofacts bot²⁰ on LINE, a popular encrypted messaging app used in Taiwan and Japan. If a user sees a suspicious story, it can paste the link in a chat to the Cofacts bot. If the story has never been seen before, a team of humans fact checks the story and uploads a message summarizing the veracity of the story to a central database. The bot then forwards this message to the original user and any other users who are curious about whether the story is true or not. Cofacts allows public access of anonymized data it has gathered on [Github](#), and also allows users to search this database through [a public website](#). Meedan, a company that supports fact checking and other online research, also has released a similar set of tools on [Check](#) that help various users automate and collectively manage workflows for the fact-checking process on WhatsApp and other platforms.

[Certain tools exist, such as Backup WhatsApp Chats](#), that allow users to export WhatsApp conversations to CSV files, but a user must still belong to a channel to export the chats from it. Telegram, an encrypted messaging app, has an API that allows users to access public channels programmatically. This allows some amount of public media monitoring, though the data is not as rich as on Twitter.

Helpful Network Visualization Tools and Packages

The table below lists several popular tools and code packages that are used for network visualization and network analysis.

Network Visualization and Analysis Tools			
Open-Source/Free Tools	Paid Tools	Commonly used Packages (Python)	Commonly used Packages (R)
Gephi NodeXL	Graphika Affinio	networkx matplotlib igraph	igraph plotrix

¹⁷ <https://www.niemanlab.org/2017/03/to-slow-the-spread-of-false-stories-on-whatsapp-this-colombian-news-site-is-enlisting-its-own-readers/>

¹⁸ <https://www.niemanlab.org/2018/06/whatsapp-is-a-black-box-for-fake-news-verificado-2018-is-making-real-progress-fixing-that/>

¹⁹ <https://www.cddwestafrica.org/whatsapp-nigeria-2019-press-release/>

²⁰ The Chinese name of this bot is 真的假的, "true or false".

Identifying Influencers, Groups and Accounts

The goal of collecting and analyzing data is to understand how information is being distributed within a network. What stories are gaining most traction? Which users are most influential? What news domains are most frequently cited in the conversation? With a strong dataset and the proper tools, you can begin to answer these questions with specificity and glean the dynamics of information dissemination in the online space you are observing.

There are two ways to conceptualize influence on social media: we can call these *content-based* and *actor-based* methods of identifying influence. We'll explore both of these below.

Content-Based Methods of Identifying Influence

Content-based methods focus on Tweets, hashtags, keywords or websites in the form of URLs or domains²¹. In some scenarios, actors worth watching will be known - outlets or politicians that frequently disseminate disinformation, for instance. On the other hand, it's quite common to not know the sources of disinformation, hate speech or other forms of content you are looking for.

Often it can be most useful when analyzing online conversations around elections or other political discourse to understand what content is gaining the most traction - especially when there isn't a particular actor you are looking to analyze that you have identified in the data. Looking at which URLs or Tweets are gaining the most traction in the data can be a good place to start in this scenario.

Tweets

When examining Tweets, influence can be approximated by looking at the number of retweets and/or likes it has garnered²². Whether you're using a third-party tool or pulling data from the Twitter API, you should have easy access to this data at all times. After determining which tweets are the most influential, you can use that Tweet as a springboard for further investigation. Some questions worth investigating include:

- Who originally produced the Tweet? Who has retweeted the post? What do the followings of these users look like? If they are large, it may be worth producing network analysis of them.
- What hashtags are used in the Tweet? If any of them are distinctive or only pushed by a small set of users, is there anything that these users have in common?
- What URLs are present in the post? If the URL is suspicious (recently created or promoting disinformation), it may be worth some extra investigation - checking the domain registration records through a Whois lookup²³, or searching Twitter for other interesting mentions of the URL, for example. You could also use the CrowdTangle browser extension to see if this URL is gaining traction on Facebook, Instagram, Reddit or elsewhere on Twitter.

These same strategies and principles also hold true for Facebook, Twitter, Gab and other social media platforms - analyzing interactions with a post is a reliable way of measuring the influence of a message in a given community.

Hashtags/Keywords

Hashtags are naturally one of the main entities of interest when examining Twitter and other social data. The convention of using a hashtag to highlight the topic of what's being tweeted is a blessing for researchers - it enables us to collect relevant conversation data for a topic of interest with great ease. Once a hashtag or set of hashtags of

²¹ URLs refer to a full link that navigate you to a particular story or site, while domains refer to the site that is hosting that content - this corresponds to the text preceding the top-level domain (abbreviated as [TLD](#) - such as [.org](#), [.com](#), [.gov](#), etc.) and the TLD itself. For instance, the three URLs [example-news-site.com/story1](#), [example-news-site.com/story2](#) and [example-news-site.com/story3](#) are all hosted on the same domain - [example-news-site.com](#).

²² An interesting note to keep in mind for Tweets is that Tweets tend to have more likes than retweets. This is similar to the fact that most Facebook posts have more likes than shares. If these ratios are abnormal, it can be an indicator of inauthentic activity, though by no means a guarantee.

²³ There are many Whois databases to check registration details, one reliable one is maintained by the Internet Corporation for Assigned Names and Numbers (ICANN) <https://lookup.icann.org/>

interest have been identified, several possibilities exist for deeper investigation: analyzing the most commonly co-occurring hashtags, splitting up hashtag citations by time, or analyzing co-occurring URLs are just a few options.

URLs/Domains

URLs and domains occurring within Tweets, most often linking to news sources, are an extremely useful source for determining what content, publications and narratives are gaining the most traction within a given online community.

A common first step URL/domain analysis is to extract unique URLs from a dataset and count the number of times they are cited within the set. This technique is simple and powerful, but it is also easy to go wrong in a number of subtle but consequential ways. To ensure your analysis yields the most accurate and useful insights, it's worth paying attention to a few of the issues below.

- **Resolving shortened URLs** - URLs in Tweets are often shortened to fit within character limits - usage of a URL resolving tool will give you the full URL that the shortened URL points to. Certain tools, such as [URLex.org](https://urlx.org/), offer an API for bulk, automated resolution of URLs. This is the best way to ensure you aren't missing data in shortened URLs.
- **Standardizing URLs** - The same domain can be cited using different strings of text²⁴. For example, links to the *New York Times* may occur in text as www.nytimes.com, nytimes.com, NYTimes.com, <http://nytimes.com>, nyti.ms, <https://nytimes.com>, <http://www.nytimes.com>, <https://www.nytimes.com>, or m.nytimes.com - these are but a few of the possibilities. In order to ensure you don't underestimate the influence of a URL or domain by counting different strings, you'll want to make sure you standardize the format before counting. It won't always be possible to control for all the diverse ways that URLs linking to the same content can occur, but a little thought given to standardization ahead of time can go a long way towards enhancing your analyses. Things to consider when standardizing are (1) case sensitivity²⁵, (2) prefixes (<http://>, <https://>, www., ww2., etc.), and (3) subdomains, among other factors.

After analyzing the most popular URLs in a set, you have several options for further analysis. If the domain is a relatively new news source on the scene, you can use open-source intelligence (OSINT) techniques such as checking the domain's registration information to start to gain insight into the domain's connections. OSINT is a useful way to gather more information about accounts or websites of interest. It is a constantly changing field, in which tools and techniques appear and disappear every day. One of the best sources for learning new skills is [Intel Techniques](https://www.bellingcat.com/), but there are numerous others. Bellingcat, a collective of researchers that often documents Russian influence operations, has made an artform of using OSINT for delivering groundbreaking journalistic stories, and maintains a public Google doc²⁶ with a thorough inventory of investigative and OSINT tools. Lawrence Alexander, a data scientist based in the UK, astutely used [Google Analytics codes](https://www.google.com/document/u/1/d/1BfLPJpRtyq4RFtHJoNpvWQjmGnyVkfE2HYolCKOGguA/edit) to map out a network of pro-Kremlin websites in 2015.

After gleaning the most influential websites and articles in a given dataset, you can also do some content analysis to analyze what narratives occur in these articles. NLP (natural language processing) techniques such as using n-gram frequencies to analyze the most common phrases in these articles, tf-idf to compare the relative themes of different articles, or use of qualitative content analysis methods can all lead to informative findings once you've narrowed down some influential articles and URLs.

²⁴ Text data is often referred to in a computational context as "strings". This is short for *strings of characters*, and is how computer scientists will often refer to the Tweet/post text in social media datasets.

²⁵ Case sensitivity refers to whether text data is lowercase or uppercase. These problems are most easily handled by lowercasing all URLs in the set before counting each URL's number of occurrences. It is important to note however that while most full URLs are not case sensitive, many shortened URLs are. It is therefore recommended that researchers first handle URL resolution before moving to lowercasing/standardizing URLs.

²⁶ <https://docs.google.com/document/u/1/d/1BfLPJpRtyq4RFtHJoNpvWQjmGnyVkfE2HYolCKOGguA/edit>

Network-Based

Another way to analyze influence in a dataset is to take a network-based approach. In this approach, you use your data to build a relevant network as discussed in the *Data and Network Analysis* section above - this could be mentions, followers, retweets, or other network based metrics.

Clustering/Groups

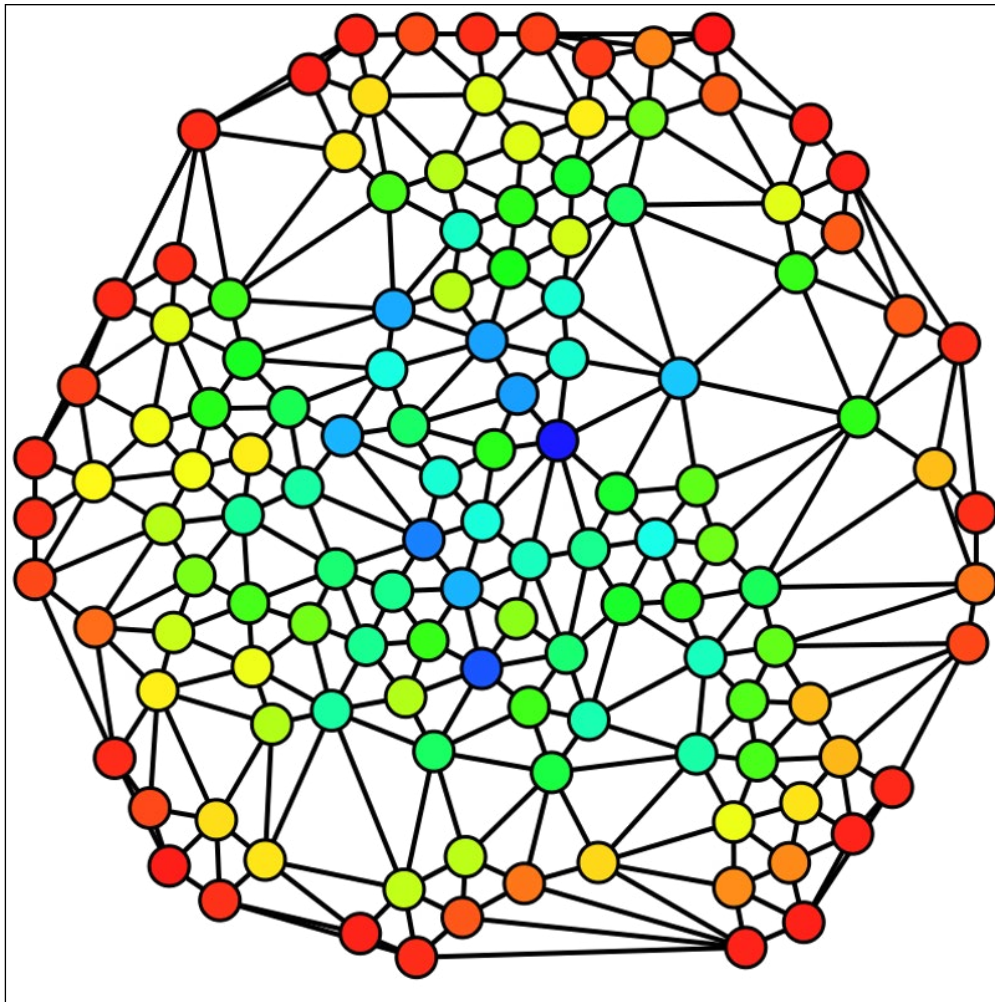
Detecting communities is the basis for most significant work that compares communities (often referred to as *clusters* in SNA). While there are complex methodological theories that underlie different algorithms for clustering, most of the hard work is done for you in any software you would use. Gephi offers several options for layout and community clustering (you can find more details in the tutorial [here](#)). Graphika also does this work for you automatically. Suffice to say, once your network visualization and analysis software determines the number of different communities in your network, you can move on to qualitative analysis to determine what members of a given cluster have in common.

Oftentimes, communities and clusters of accounts will have discernible features in common, such as promoting similar news sources or belonging to a similar party. This is where understanding the overall political context becomes important. Combining quantitative analysis of the community's data with qualitative content analysis is the best way of determining what members of a given community have in common.

Centrality

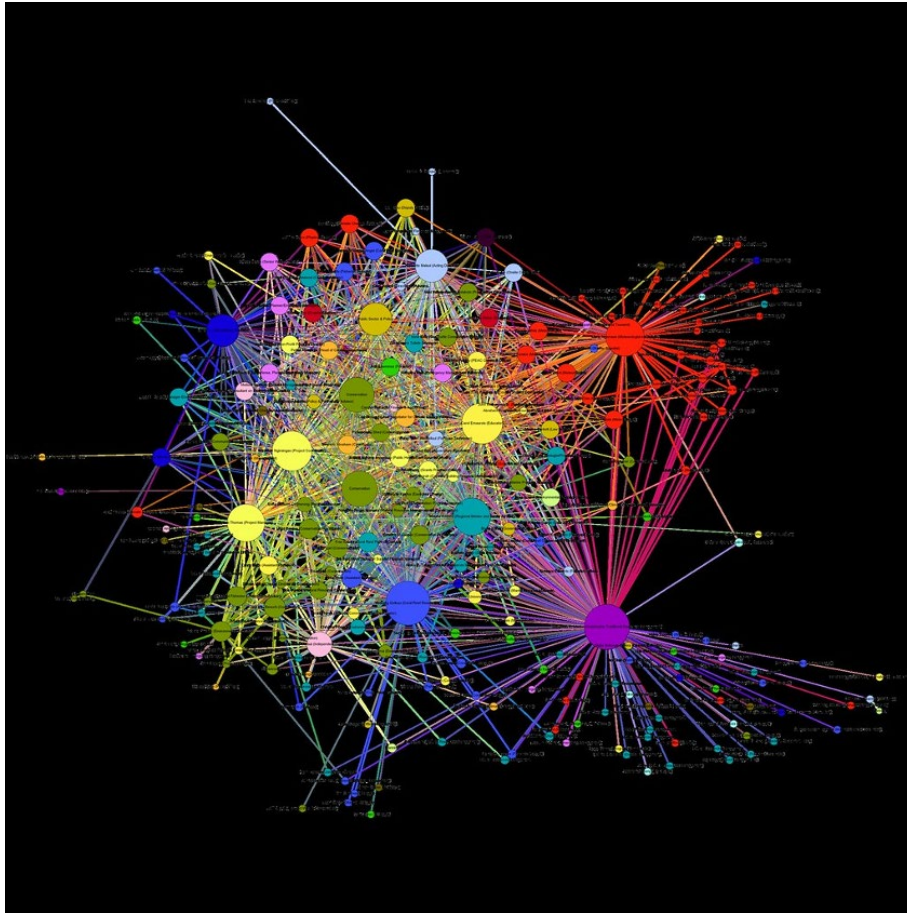
In network theory, specifically as it applies to SNA, the main metric for analyzing the influence of a single actor or group of actors is *centrality*. Centrality is a metric for how well-connected, or how influential, a given node is within a network. There are several different methods for measuring centrality.

- **Degree centrality** - In an undirected graph, the degree centrality of a given node is the number of direct connections it has to other nodes around it. In a directed graph, where the direction of the edge is taken into account, there are two types of degree centrality - *in-degree centrality*, or how many incoming connections a node has - and *out-degree centrality*, or how many outgoing connections a node has. In-degree centrality in these scenarios is arguably most relevant - a node with a high number of incoming connections has high potential for influence. This is a relatively intuitive concept - an account with 15,000 followers (an in-degree of 15,000) has a greater ability to directly influence than an account with 100 followers. The importance of in-degree/out-degree will vary depending on the network you are examining.
- **Betweenness centrality** - Betweenness centrality can be thought of as measuring the capacity of a node to spread a message to other nodes quickly. Put another way, betweenness centrality gives higher scores to nodes that are likely to be vectors of virality or quick message dissemination. This value is measured by counting the number of times a given node finds itself in the as part of the shortest path from one node to another node.



Hue (from red=0 to blue=max) indicates each node's betweenness centrality. (Claudio Rocchini) [Creative Commons BY 2.5](https://en.wikipedia.org/wiki/Social_network_analysis#/media/File:Graph_betweenness.svg)
https://en.wikipedia.org/wiki/Social_network_analysis#/media/File:Graph_betweenness.svg

- **Eigenvector centrality** - Eigenvector centrality is very useful for understanding influence within a given network or sub-network. The idea behind eigenvector centrality is essentially “Who influences the influencers?”. With eigenvector centrality, influence is less a matter of the number of ties a target node has to other nodes, and more a matter of how many connections the target node has to other well-connected nodes. You can think of this version of centrality as a type of “trickle-down influence” - how likely is it that information flowing from a source node spreads to other influential nodes? Google’s search engine uses a form of this concept as it ranks pages (or nodes) according to the number of other pages that reference them. Nodes with high eigenvector centrality can play a particularly powerful role in the dissemination of goods (such as money, information, germs, etc.) through a network.



Pacific RISA Social Network Analysis Project, Palau-centric map; all members are from or connected to Palau. Nodes are sized according to eigenvector centrality and colored according to profession. [Creative Commons BY 2.5 https://www.flickr.com/photos/pacificrisa/11344578486](https://www.flickr.com/photos/pacificrisa/11344578486)

- **Bridges** - While not a form of centrality, *bridges* are a useful network structure phenomenon to keep in mind. A node or small set of nodes act as a *bridge* when they connect one dense cluster to another dense cluster. Without bridges, clusters of diverse interests (such as members of two separate political parties, or political aficionados in two different countries) would have no means of transmitting information or communication between each other. In this regard, bridges are important parts of viral content on online networks.
- **Leaves** - *Leaves* are nodes in a graph with only one edge connecting them to a network. These accounts are often worth removing from a network, especially large networks, in which they are highly unlikely to have any influence.

Several useful online tutorials and resources are available for reading about different types of centrality - one helpful resource is the book *Analyzing Social Networks* ([Borgatti et al., 2013](#)), which lays out different types of centrality and how they are measured in Chapter 10.

Open-source tools like Gephi can do the difficult math of calculating different forms of centrality for you in seconds²⁷. The choice of which form of centrality to use is ultimately a choice that lies with you and your team. While consulting a data scientist or network theorist is ideal for making that choice, you don't need to stress too much about the methodological minutiae here - nearly all forms of centrality assume that the most influential nodes are, in some sense, the nodes with the greatest number of ties in a network or subnetwork.

Once you've determined which form of centrality you'll be interested in using, you can even customize the visuals of your network to reflect this. Gephi allows you to [use node size or color shades to highlight centrality](#), for instance.

²⁷ Or minutes, depending on how large the network in question is.

Identifying News Sources

For elections and information analysis in particular, the question of how many news sources are cited in the data is of central importance. This is an area in which it is especially useful to work with individuals that have a thorough social and political understanding of the country or region in question. The most frequently cited news sources are often common knowledge to these individuals. It is important to have a strong understanding of the underlying media environment in the country, as reviewed above, including television, radio, newspaper and online media market share and influence.

There are several useful methods for identifying new news sources in a region or dataset. One option is to run frequent manual Google searches of news and politics in the region - especially in multiple languages if the region is multilingual. News sources vying for influence are likely to have strong search engine optimization (SEO) to surface them within the first results retained by Google, and can often surface recently created news sources of which the average citizen is unaware. Another option, which is highly recommended, is to perform domain and URL citation analysis on a dataset. This can take a few forms - monitoring CrowdTangle for the leading URLs in a given set of political pages, or extracting unique URLs/domains from post data²⁸ in a collected dataset with a local expert and searching for domains neither of you have seen before. Crowdtangle has [a free plugin for Chrome](#) and other browsers that allows you to see shares for an individual article, but to access their more advanced dashboard and other tools users must obtain an institutional license, often through Facebook, which now owns it.

The guiding insight here is that often disinformation emanates from newly created domains that appear quickly around political issues and elections, and just as quickly disappear. This was the case with *streetnews[.]one*, a domain that was created and promoted on Gab in the lead up to the 2018 US midterms. The domain was used to spread Islamophobic and sensationalist content before it quickly disappeared. This was also the case in [Endless Mayfly](#), an analysis of Iranian disinformation conducted by the University of Toronto's Citizen Lab in tandem with industry experts. In this case, Citizen Lab's team coined the phrase, "*ephemeral disinformation*", to describe quickly appearing and disappearing disinformation domains targeting key issues, elections and campaigns.

Using assembled lists of disinformation domains or fake news websites at this stage can be extremely helpful. Researchers at Stanford assembled a list of over 600 domains known to produce false content in late 2018, principally targeting the United States (Allcott et al, 2018)²⁹. Two important caveats should be kept in mind if you ever decide to use such lists: (1) disinformation and the online sphere move quickly - new disinformation domains that have occurred since the compilation time of the list will be missing from analysis; and (2) any list used should be properly vetted for methodological rigor - using randomly assembled lists found online could be detrimental to the integrity of the research.

Analyzing Accounts and Content

Once you have collected data and/or built a network of relevant accounts you are interested in, you and your team are ready to begin analyzing accounts and content in the dataset. This is likely to be the portion of the work on which you'll spend the most time, especially if you do not initially know what you are looking for.

Content analysis is best carried out by oscillating between qualitative and quantitative methods, often iterating the process several times to find specific accounts or content of interest. In this section, we will provide you with a few tips and tools to help guide this work.

²⁸ Preferably along with citation counts as an approximation for these sources' popularity. Other methods, such as extracting URLs from the most retweeted or most liked posts in a dataset, are also ways of approximating the influence of a particular domain within a dataset.

²⁹ You can find this paper [here](#).

Types of Content

It is useful before diving into the data to have a grasp of some of the types of content you may want to look for in the dataset.

- **Disinformation/Misinformation** - False and misleading political content is one of the most pernicious forms of content you'll want to watch for in your data set. Though you'll hear false content referred to by several names, including computational propaganda and fake news, this kind of content is most commonly referred to as disinformation or misinformation. The technical difference between these two terms lies in the intent behind the false content³⁰. The definitions we'll use to differentiate these two words are taken from Data and Society's Lexicon of Lies report (Jack 2017), which explores a set of words that are useful to understand and use when discussing false content online.
 - **Disinformation** - False content that is spread with deliberate intent to deceive. Politically motivated nation-state actors or financially motivated actors are the most likely actors to knowingly distribute disinformation, as defined in this way.
 - **Misinformation** - False information that is unknowingly spread and distributed. If an account promotes a story without having an intent to deceive users, this qualifies as misinformation.
 - **Malinformation** - Certain authors have also highlighted the phenomenon of "malinformation" - the distribution of true or mostly truthful information with intent to harm. Claire Wardle and Hossein Derakhshan define malinformation as "*Information that is based on reality, used to inflict harm on a person, organization or country*" (Wardle and Derakhshan 2017).
- **Hate Speech** - hate speech is language that demonizes people of a target race, ethnicity, gender, sexual orientation or religion. Often, hate speech incites others to harass, denigrate or even engage in violence against the targeted social group.

Linguistic Analysis

One way to go about analyzing content in a rigorous way is to conduct linguistic analysis. Linguistic analysis can help you understand the main languages of messaging, main content themes, narratives being pushed and developed, new keywords and jargon, and whether hate speech or other dangerous content is occurring in the conversation.

Keywords and Lexicons

Not every post or tweet in a given dataset is likely to be relevant to the specific questions in which you are interested. For that reason, it can be useful to compile a list of relevant keywords that are likely to contain information relevant to your inquiry. Working with a subject matter expert - someone who has a deep understanding of the language and politics of the region of interest - is the best way to ensure the quality of a keyword list. These keyword lists are also sometimes referred to as "lexicons" if they all relate to a common theme. For instance, lexicons of hate speech in different languages that are related to different political contexts are often used to analyze political content. Even a quick hour session with a subject matter expert can go a long way towards introducing some rigor into the compilation of keyword lists to ensure quality. You can also consult relevant academic literature and use previously compiled keywords lists if they are of high quality and relevant to your context.

Examples of Lexicons

There are ample examples of lexicons that are available online and useful for linguistic analysis and relevant post extraction. NDI's Gender, Women and Democracy team compiled [a hate speech lexicon](#) relating to Indonesian, Kenyan and Serbian languages (Zeiter et al., 2019). PeaceTech Lab, a non-profit organization dedicated to using technology to promote peace in developing countries, also has several publicly available hate speech lexicons

³⁰ While this is the case, it is often not possible to know the intent behind false content that spreads online. For this reason, you are likely to hear these words used interchangeably at times.

available for free on its website. These lexicons are multilingual, which can be extremely useful for monitoring conversations in multiple locations or in regions with high linguistic diversity.

Hatebase.org also has multilingual hate speech keyword lists. Scholars Roya Pakzad and Nilhoufar Salehi used a list compiled from this website for [their study of computational propaganda targeting Muslim Americans](#) in the 2018 US midterms (2019). A similar list was used for [a quantitative analysis of Islamophobia on Gab](#) in the lead-up to the same elections (Woolley, Pakzad & Monaco, 2019).

Qualitative Linguistic Analysis

Once you have a relevant set of posts extracted, you can dive into linguistic analysis. Qualitative methods are those that analyze the content and messaging themes of posts in your dataset. This kind of work is most effective when done by humans and carried out by an expert or team of experts that is familiar with the political and linguistic context of the region in question.

Whether using qualitative methods, quantitative methods or both during your analysis, it is most important that your methods are consistent and systematic. Using the same methods to analyze every post or part of your dataset is the best way to guard against introducing bias into your results.

Narrative Analysis

Narrative analysis is an intensive way to analyze linguistic content in a dataset, but it can be very interesting. Narrative analysis analyzing how certain outlets refer to a given topic, especially over time, can shed a light on framing and influence of outlets on public opinion.

Qualitative Coding

Another qualitative method that is useful for understanding what sorts of messaging are occurring in a given dataset is referred to as qualitative *coding*. This consists of a team of experts independently assigning one of a set of predefined categories (or “codes”) to posts one-by-one. After categories are assigned to each tweet, quantitative analysis can be carried out.

For example, in a hypothetical study on a presidential election between two candidates - candidate A and candidate B - in the imaginary country of Qumar³¹, we can imagine five possible categories for messages collected around the election:

1. Pro-candidate A
2. Pro-candidate B
3. Anti-candidate A
4. Anti-candidate B
5. Neutral

After a team of experts has coded all the posts in the data set, there are several quantitative analyses we could undertake as a next step:

- **Distribution of posts in each category** - We could analyze the amount of posts that occur in each category to examine whether there was greater support or opposition online for candidate A or candidate B.
- **Keyword usage by category** - We could also examine posts from each category for relevant keywords, to see whether supporters or opposition of either candidate use certain words.
- **Bot content in each category** - Another possibility is reviewing posts from each category for bot content to see if either candidate is garnering greater support or opposition from automated agents online.

³¹ Qumar is a fictional country that appears in the US political drama series *The West Wing*.

These are just a few examples of analyses that can be undertaken after high-quality qualitative coding is undertaken.

Quantitative Linguistic Analysis

Quantitative linguistic analysis is also known as natural language processing (NLP) or computational linguistics. The methods used in quantitative linguistics analysis stem from the idea that we can gain certain insights about the types and frequencies of certain messaging when we view language from a statistical standpoint. Put more concretely, if we aggregate words or sets of words from relevant posts in our dataset, we can examine what sorts of themes emerge. This subsection will introduce you to some of the basics of extracting word counts from a set of data to examine common themes online.

While the techniques in this section can be implemented with any programming language, it is worth noting that Python and R come with several packages that simplify the process. These packages, such as Python's Natural Language Toolkit ([NLTK](#)) or R's [tm](#) package (tm standing for *text mining*) have extensive documentation in the form of books, online guides and tutorials that can teach you the basics within a few hours. We highly recommend taking some time to familiarize yourself with one of these packages, especially if you already know a bit of Python or R. A few extra hours upfront will save you and your team lots of time down the road!

N-grams

N-grams are the building block of nearly all quantitative linguistic analysis of social media content. An *n-gram* is a sequence of words of length *n*. Three types of *n-grams* in particular are most common for quantitative linguistic analysis of texts:

- **Unigrams** - A single word can also be thought of as a sequence of 1 word. This type of *n-gram* is known as a *unigram*.
- **Bigrams** - A pair of words occurring *next to each other* in a sentence form a bigram. Put another way, a bigram is any sequence of 2 words that occurs in a text.
- **Trigrams** - By now, you have probably guessed that a trigram is any triplet of words that occurs in a sentence. Every sequence of 3 words within a single sentence or text forms a separate trigram.

It is most common to work with these three types of *n-grams*. This is largely due to the fact that using larger values for *n* is "computationally expensive" - you are likely to slow down the performance of your computer and not get much bang for your buck in return. For sequences of words greater than 3, it is common convention to simply refer to these the number in question followed by the word "-gram". For instance, 4-grams would be a sequence of 4 words, 5-grams would be a sequence of 5 words, etc.

The key insight behind *n-grams* is that you not only have word counts, but you also have a small amount of the *context* a word occurs in. Knowing that you have 16 occurrences of the word "king" tells you that you have a frequent topic on your hand, but not much more than that. If you expand your analysis to 4-grams and see that you have 15 occurrences of "down with the king" and one occurrence of "long live the king", you can say with certainty that most "king"-related posts in your dataset are not messages of support.

Once your team is comfortable with the technique of extracting *n-grams* and counting *n-gram* frequencies within a text, you can apply this technique to different parts of your dataset. For example, comparing *n-grams* among accounts or pages that support different political candidates/parties may be enlightening. Comparing *n-gram* frequencies from different media outlets producing articles relevant to the same election may lend insight into the main focus of each outlet. *N-gram* frequency comparisons between post histories of two separate accounts or pages may elucidate their favorite messaging topics. These are just a few examples of possibilities that could help enhance your research.

Other Quantitative Linguistic Techniques

Assembling n-gram frequencies is a technique that can be applied as a first step to other NLP techniques that can help analyze a dataset in greater detail. For example, after extracting word frequencies or n-gram frequencies, a statistical technique known as *term frequency-inverse document frequency* (tf-idf) can be used to compare the messaging themes of different documents relative to each other. These “documents” could be collections of articles from different news outlets leading up to the election or post histories for different accounts of interest, for instance. Tf-idf is a basic way of determining what unique themes distinguish one document from another. This technique could be used to analyze which accounts most frequently promote a given party, or what makes one community’s messages distinctive from all the others you’re examining.

A quick primer on how to use tf-idf for linguistic analysis³² and content clustering can be found in *Mining the Social Web* (Russell and Klassen 2018). This discussion is well worth consulting - it includes example code in Python and is particularly easy to follow.

Automated Hate Speech Detection

To analyze hate speech in a dataset, you will most likely first have to compile a list of sensitive hate speech terms that are relevant to the region you are examining. These terms should be reviewed with a subject matter expert in detail and in every language likely to be used in the region in question, methodology that is described in the NDI report “Tweets That Chill”. (Zeiter et al., 2019)

One heuristic for automatic detection of hate speech without keyword compilation is using Google Jigsaw’s Perspectives API - an open-source tool that was introduced in 2017. The Perspectives API currently only provides support for English language comments³³. It takes a string of text as input and outputs a toxicity score for that text. The higher the score, the more “toxic” the utterance was judged to be by the Perspectives API’s machine learning models. An example of usage of this API can be found in Pushshift.io’s Gab dataset. In addition to containing posts from Gab, Pushshift.io ran each post through the Perspectives API and recorded the toxicity score it received.

The Perspectives API is currently one of the only publicly available tools that assigns toxicity scores to language. It takes great time and effort to capture the nuances of context and intention in human language, and for this reason, sentiment analysis and hate speech detection tools are still in their infancy. On top of the inherent general difficulty of the problem, the diverse specificity of the context of each language and social context in which hate speech occurs compounds the difficulty of producing rigorous and reliable automated hate speech detection tools. For these reasons, working hand-in-hand with subject matter experts to assemble relevant keywords and lexicons is still the best method for analyzing hate speech within a social media dataset.

Bots

Much has been written on the influence of bots on social media in the past few years. The Computational Propaganda Project (ComProp) at the University of Washington and the Oxford Internet Institute (OII) did pioneering work on bots, exploring their influence in disseminating and promoting computational propaganda in countries around the world.

Bots, simply put, are computer programs that control profiles on social media sites, often posing as real people and interacting with other humans online. In the case of social media, most bots are controlled through computer

³² This text will also introduce you to some relevant techniques for enhancing your analysis, such as stopword removal and stemming.

³³ English has disproportionately benefited from the attention of linguists, both computational and non-computational, throughout the 20th and 21st centuries. Languages which have not benefited from great attention and research are often referred to as *low-resource languages* in linguistics and NLP. The importance of producing research on these languages cannot be overstated. If you or your team undertake hate speech or linguistic research as part of your media monitoring efforts, it may be worth reaching out to professional linguists in industry or academia to see whether your research could help contribute valuable insights to the field.

programs that control accounts through the API³⁴. The simplest of these bots often have telltale signs that give them away as being automated - tweeting at the same minute of every hour or not having a profile photo, for instance. These pieces of data are often used in machine learning algorithms for tools that tell bots from humans online.

Tools for Detecting Bots

While there is most often no surefire way of saying with certainty whether an account is a bot or not, there are useful options available to the curious researcher for detecting bots online. One of the best options currently available is Botometer, a tool that uses machine learning to assign a bot probability score to input accounts. Botometer has a long history of academic research and development at the University of Indiana and is free to use. A number of other classifiers also exist (see table below).

Bot Detection Tools

Tool Name	Can be implemented through code for classification of batches of accounts?	Platform ³⁵	Extensions / Website available?
Botometer	Yes (Python)	Twitter	Single accounts can be checked on the website
Tweetbotornot	Yes (R)	Twitter	Only code implementation available.
Botcheck.me	No	Twitter	Chrome and Firefox Extensions, last updated in 2019.
Botsentinel	No	Twitter	Android App, and Chrome/Firefox extension
Pegabot	No	Twitter	Website

³⁴ In the data collection section above, we explored the differences between APIs and web scraping - this difference is also useful for understanding how certain bots operate on social media and on the web at large without using APIs. Bots can be programmed to interact with everyday webpages - in fact, this is quite trivial for programmers to do. For example, bots can go from site to site and analyze the content of web pages - these types of bots are often called *crawlers* or *spiders*, and in fact, this is how Google gathers data on web pages to put in its search engines. Just as not all bots are social media bots, not all bots are bad, and in fact some of them are necessary for the internet as we know it to function every day.

³⁵ Open-source bot detection tools have only been developed on Twitter - largely because the Twitter API offers a rich set of information on users (public metadata), which are useful as features in constructing machine learning models that classify accounts as bots or humans. Analyzing bot activity on other platforms largely relies largely on manual analysis and investigation, such as spotting superhuman activity (100 posts/min, posting messages in regular intervals, etc.). For instance, in a [study of Islamophobia preceding the 2018 US midterm elections](#) on the social media platform Gab, we detected the presence of a disinformation bot by analyzing identical posts occurring in short intervals emanating from one user.

Conclusion

Ultimately, all of these techniques and tools should help form the user's approach to data collection and analysis. Users should also consider building a workflow using these techniques, a system of archiving and sorting information collected, and potentially reporting it to platforms or other national agencies that can act on the data collected. Depending on the subject, they should consider a number of the different tools and techniques described above. Whether searching for disinformation during an election or hate speech and computational propaganda in traditional online political discourse, these resources can be applied in different ways. Teams and researchers should also consider the resources they have - human, financial, and technical - as they construct their project.

Many of the tools mentioned are open-source, but others are expensive and often complex to apply without sophisticated knowledge of the tools and methods involved in developing them. Often it is worth considering simpler solutions that can be applied by less experienced researchers, or working collaboratively across teams with different skill sets to unearth different insights in the same dataset. Consider partnering local researchers with international technical experts to gain new kinds of insights, and work to build methods for collaboration online across communities, countries and regions.

One way to find partners with whom to collaborate, in your local context and internationally, and to link with social media companies for reporting and research is through the Design 4 Democracy Coalition. The D4D Coalition is a group of International NGOs including NDI, the International Republican Institute, the International Foundation for Election Systems, International IDEA, and national civil society organizations all over the world engaging with technology companies such as Facebook, Microsoft and Twitter to encourage them to design their systems, content moderation and policies for democratic principles. Research and online monitoring have become crucial components of elections and democracies the world over, and this guide is written to give groups working to support these ideas through policy and technical systems the tools, methods and capabilities to do this work and help better inform society.

See below for further references, open-source tools and examples of code and a walkthrough for using Twitter's API.

Appendix I: Example API code - Collecting Data from Twitter's Search and Stream APIs with the Rtweet Package

In this section, we'll briefly introduce you to some code that you can use to gather historical and live Twitter data. With just a few lines of code, you can easily collect Twitter data relevant to your election context. After gathering the target data, you can export the desired data to CSV, proceed to analyze it in Excel or Google Sheets, or pass this data onto a dedicated data scientist on your team to search for deeper insights. Being able to collect data on your own in real time, even if you may not have a dedicated data scientist on your team at the time of collection, is extremely useful. Counterintuitively, data often becomes *more* valuable over time because it captures disinformation, bots and nefarious actors whose actions and posts may later be removed from the platform for violating terms of service.

Step 1: Download R, RStudio and Rtweet

In this guide, we'll be teaching you to pull some basic data from the Twitter API using the programming language R. There is a specific package in R, called *rtweet*, that makes pulling data from Twitter extremely easy.

You will have to download and install a few items to get up and running.

- Rstudio: <https://www.rstudio.com/products/rstudio/download/>
- R: <https://www.r-project.org/>

Step 2: Apply for a Twitter application, retrieve your application keys and create your token

In order to collect data from Twitter, you use an *application* - this is a secure means of using your account to interact with the platform through code. In years past, Twitter applications were free and no approval was needed - nowadays, you have to apply for an application, which you can do [here](#).

Once you've retrieved your application, you'll want to log into Twitter on a browser and head to <https://apps.twitter.com>. You can retrieve your application's consumer and access keys here - there will be four total keys. These keys are simply strings of text that your application will use to verify that it is indeed requesting data on behalf of a person - in this case, you. This usage of keys for *authentication* is a process known as *oauth*, which stands for open authorization. It developed as a way to allow applications to perform actions online on behalf of users without transferring their password and username for every action.

Once you have your keys and your app name, you can use the code that follows in the screenshot below to create your authorization "token". Once this token is created, you are ready to start using the Twitter API.

```
library(rtweet)
app<-"<app name here>"
consumer_key<-"<consumer key here>"
consumer_key_secret<-"<consumer key secret here>"
access_token<-"<access token here>"
access_token_secret<-"<access token secret here>"
create_token(app,consumer_key, consumer_key_secret, access_token, access_token_secret)
```

The lines of R code above are initial lines you can use to authenticate your Twitter application and connect to the Twitter API. Once you've applied for access to and been approved for a Twitter application, you can retrieve your consumer and access tokens from Twitter. You'll use these tokens to authenticate your application (shown in the code above). Once your application is authenticated (i.e. Twitter knows that the application is pulling data for you, and not for someone else), you are ready to begin interacting with and pulling data from the Twitter API.

University of Missouri Professor and *rtweet* developer Dr. Mike Kearney also details the authentication process step-by-step on the official website for *rtweet* [here](#).

Gathering Historical Data:

Using `rtweet`, you can easily gather Tweets using one or several hashtags of interest from Twitter's Search API. This API is *historical*, which means you'll be gathering historical data from the past 7-9 days that corresponds to your query. Queries are simply criteria you are interested in having in the Tweets you gather. Queries can contain hashtags, keywords, account names, URL, a combination of these, or all four. In this section, we'll focus on hashtags, but the process and syntax are exactly the same for other entities.

The main function you'll use for searching for historical tweets is called `search_tweets()`.

```
my_data<-search_tweets("#ExampleHashtag", n=50000, retryonratelimit=TRUE)
```

This line of R code queries the Twitter Search API for up to 50,000 Tweets that use #ExampleHashtag from the last 7-9 days, and saves the results in a dataframe called "my_data". If the hashtag was used fewer than 50,000 times in that timeframe, a smaller number of Tweets will be returned. Changing the input value for "n" can increase or decrease the maximum amount of Tweets you will retrieve.

```
multiple_hashtags<-search_tweets("#ExampleHashtag1 OR #ExampleHashtag2", n=50000, retryonratelimit=TRUE)
```

This line of R code is similar to the query above, but returns up to 50,000 Tweets containing #ExampleHashtag1 or #ExampleHashtag2. This syntax can be used to query Twitter's Search API for multiple hashtags. In this example code, the results are saved in a dataframe called "multiple_hashtags". Multiple hashtags are separated by what are called "boolean operators": these are simply "AND" or "OR". Use "AND" if you only want Tweets that contain both hashtags; "OR" will return tweets containing either.

Streaming Twitter Data in Real Time:

You also have the option of streaming Twitter data in real time. This query requires you to specify the length of time you'd like to stream Tweets in seconds, as well as the entities you'd like to stream (hashtags, URLs, keywords, @-mentions, etc.).

The syntax for streaming tweets in `rtweet` is slightly different than that used for querying the search API. When streaming Tweets, you'll want to use the `stream_tweets()` function. Multiple query items will be separated by commas.

```
my_streamed_data<-stream_tweets(q="#ExampleHashtag1", timeout=60)
```

```
my_streamed_data_w_multiple_hashtags<-stream_tweets(q="#ExampleHashtag1,#ExampleHashtag2", timeout=60)
```

The figures above show how to use `rtweet`'s `stream_tweets()` function to gather Tweets in real time. The upper figure collects any Tweets using #ExampleHashtag1 during a 60 second stream window. The bottom figure does the same, but also collects Tweets containing #ExampleHashtag2.

Writing Results to a CSV Output File:

A commonly used file format for data analysis is the CSV file, which stands for "comma-separated values". In a CSV, each line represents a single row of the spreadsheet³⁶, and each entity between commas represents a cell - or more precisely, a value³⁷ within a cell.

A CSV file is essentially a spreadsheet that is in a machine-readable format. Once you have a CSV of the Tweets you've collected, you can pass it on to a specialist data scientist for deeper analysis, or load it into a spreadsheet processor such as Microsoft Excel or Google Sheets to do some analyses on your own.

³⁶ These rows are also sometimes referred to with a few other terms by data scientists: *record*, *instance* and *observation* are other terms you are likely to hear in this context, which all are synonymous with "row" when referring to a csv.

³⁷ *Values* in a csv or spreadsheet can also sometimes be referred to as *fields*.

```
write_as_csv(my_data, "my_data_as_a_csv_file.csv")
```

This line of R code uses the *rtweet* package to write a dataframe of Tweets called “my_data” to a CSV file called “my_data_as_a_csv_file.csv”. After exporting your Tweets to a CSV file, your data can be easily shared or imported into Microsoft Excel or Google Sheets.

Closing Thoughts

Being able to pull data from Twitter and write it into a CSV is an invaluable skill. Even without programming experience, anyone can learn the process we describe here in under 2 hours, and likely even more quickly. Once you have the ability to retrieve and store CSVs of relevant Twitter data, you are able to put yourself and your team in a position to analyze valuable data now and in the future.

If you’re storing this data long-term, you may also want to consider compressing the file into a .zip format (or another format, such as a Tarball). This can reduce the amount of storage space it takes to store the file, and it makes it easier to distribute the file to other people and devices.

Appendix II: OSINT Tools

Open-source intelligence, commonly referred to as OSINT, is the art of investigating a question using only publicly available (or “open-source”) information and data. In the context of disinformation and social media monitoring, OSINT can often provide additional details on nefarious or suspicious actors online, including fake accounts or disinformation websites. Below is a list of some valuable resources for learning OSINT techniques.

- Michael Bazzell’s website and book:
 - <https://inteltechniques.com>
 - [Open-Source Intelligence Techniques](#)
- Bellingcat’s Online Investigation Toolkit: <https://docs.google.com/document/d/1BfLPJpRtyq4RFtHJoNpvWQjmGnyVkfE2HYolCKOGguA/edit>
- BuzzFeed journalist Craig Silverman’s public list of verification and OSINT tools: <https://docs.google.com/document/d/1ZJbIUk5L8fe3VKK9CLVNMj9qOFdXG-RhQT6pyEgsS4I/edit>
- Comprop Navigator, published by the Oxford Internet Institute’s Computational Propaganda project, compiles methods and OSINT tools related to disinfo and other online research. <https://navigator.oii.ox.ac.uk/>
- First Draft Guide on NewsGathering and Monitoring on the Social Web https://firstdraftnews.org/wp-content/uploads/2019/10/Newsgathering_and_Monitoring_Digital_AW3.pdf?x36710
- Fighting Disinformation Online: A Database of Web Tools, hosted by the Rand Corporation. <https://www.rand.org/research/projects/truth-decay/fighting-disinformation.html>
- Verification Handbook for Disinformation and Media Manipulation <https://datajournalism.com/read/handbook/verification-3/>

References

- Allcott, H., Gentzkow, M., & Yu, C. (2018). Trends in the Diffusion of Misinformation on Social Media. <https://web.stanford.edu/~gentzkow/research/fake-news-trends.pdf>
- Borgatti, S., Everett, G., & Johnson, J. (2013). *Analyzing Social Networks*.
- Democracy Reporting International. (October, 2019) *Guide for Civil Society on Monitoring Social Media During Elections*. <https://democracy-reporting.org/wp-content/uploads/2019/10/social-media-DEF.pdf>
- Jack, C. (2017). *Lexicon of Lies*. Data & Society. <https://datasociety.net/output/lexicon-of-lies/>
- Monaco, N. (2019). *Welcome to the Party: A Data Analysis of Chinese Information Operations*. Retrieved from <https://medium.com/digintel/welcome-to-the-party-a-data-analysis-of-chinese-information-operations-6d48ee186939>
- National Democratic Institute. (May, 2019) *Disinformation and Electoral Integrity: A Guidance Document for NDI Elections Programs*. <https://www.ndi.org/publications/disinformation-and-electoral-integrity-guidance-document-ndi-elections-programs>
- National Democratic Institute. (December, 2018). *Supporting Information Integrity and Civil Political Discourse*. <https://www.ndi.org/publications/supporting-information-integrity-and-civil-political-discourse>
- Pakzad, R., & Salehi, Ni. (2019). *Anti-Muslim Americans: Computational Propaganda in the United States*. Institute from the Future. Retrieved from http://www.iftf.org/fileadmin/user_upload/downloads/ourwork/IFTF_Anti-Muslim_comp.prop_W_05.07.19.pdf
- Russell, M., & Klassen, Mikhail. (2018). *Mining the social web*. Sebastopol, CA: O'Reilly Media.
- Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe. <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>
- Woolley, S., Pakzad, R., & Monaco, N. (2019). *Incubating Hate: Islamophobia and Gab*. German Marshall Fund. <http://www.gmfus.org/sites/default/files/publications/pdf/Incubating%20Hate%20-%20Islamophobia%20and%20Gab.pdf>
- Woolley, S., & Howard, P. (2017). *Computational Propaganda Worldwide: Executive Summary*. Working Paper. 2017.11. Oxford, UK: Project on Computational Propaganda. <http://blogs.oii.ox.ac.uk/politicalbots/wp-content/uploads/sites/89/2017/06/Casestudies-ExecutiveSummary.pdf>
- Zeiter, K., Pepera, S., Middlehurst, M., Ruths, D. (2019). *Tweets That Chill: Analyzing Online Violence Against Women in Politics*. National Democratic Institute. <https://www.ndi.org/tweets-that-chill>



NATIONAL
DEMOCRATIC
INSTITUTE

NDI.ORG