1 Chapter 6: Collecting Incident Data

1.1 Data inputs: Alerts and Canaries

We receive alerts about possible disinformation incidents from members of the disinformation team, and from other teams connected to us. Typically we get alerts around an artefact or theme, e.g.

- A new narrative emerging online, either in general social media or known conspiracy / extremist / target etc groups
- A local or world event that might spark a disinformation incident
- Anomalous or significant-sized online activity that might be associated with a disinformation incident
- Command signals from known disinformation groups (e.g. qanon)

The types of artefact that we typically receive include:

- Images
- Messages, e.g. tweets, facebook posts, SMS or Messenger/Telegram etc messages
- URLs

The processes for investigating these are discussed in more depth in the next chapter.

Several accounts and groups are either known producers or early adopters of many disinformation campaigns. We've dubbed these "canaries", as in the entities that give the first signals that something is happening (canary, as in "canary in a coal mine").

1.2 Data sources: disinformation data streams

When we get our first data inputs, it's a good idea to check them against other disinformation and related data collections, to see if they've been picked up by other researchers, or those researchers have already collected data related to these inputs that can be of use to our investigation. The data feeds are continually updated, so are a good source for breaking data; the static data collections are good for finding history on data, source, narratives etc.

1.2.1 covid19-related disinformation data feeds

Narratives

- Wikipedia list of Covid19 rumours
- WHO Covid19 myths list - narratives
- EuVsDisinfo database https://euvsdisinfo.eu/disinformation-cases/
- Ryerson Claimwatch dashboard
- CMU IDEAS Center list of Covid19 disinformation narratives (click dates)
- Indiana Hoaxy (twitter, articles)

Data

- Botsentinel: lists "trollbots" (bot-like and troll-like accounts) and the themes they're promoting https://botsentinel.com/ (not just Covid19)
- Hamilton68 - live feed from accounts attributable to Russia or China (may or might not contain propaganda; useful for seeing current themes). Public version is live feeds from official Russian sites (embassies, RT etc), not trolls. Academics can ask for a more detailed feed. https://securingdemocracy.gmfus.org/hamilton-dashboard/ (not just Covid19)
- Ryerson University covid19 misinformation portal: https://covid19misinfo.org/
    - Botswatch dashboard https://covid19misinfo.org/botswatch/
- Uni Arkansas COSMOS Covid19 list http://cosmos.ualr.edu/misinformation
- Indiana University OSOME Decahose
- Facebook datafeed: Enabling study of the public conversation in a time of crisis

Domains

- Coronavirus Misinformation Tracking Center – NewsGuard

1.2.2 Covid19-related counter-disinformation feeds

- Ryerson University covid19 misinformation portal: https://covid19misinfo.org/
- Snopes: https://www.snopes.com/
- WHO COVID-19 site: https://www.who.int/health-topics/coronavirus
- WHO information network for epidemics https://www.who.int/teams/risk-communication
- Coronavirus Tech Handbook https://coronavirustechhandbook.com/misinformation
- Experts list https://twitter.com/jeffjarvis/status/1254038157244456961

- Maryland Covid19 rumour control https://govstatus.egov.com/md-coronavirus-rumor-control

1.2.3 Covid19 general data feeds

- https://crisisnlp.qcri.org/covid19 - GeoCov19 dataset of covid19 tweets (up to about 3 weeks ago; still collecting)

1.2.4 General disinformation datasets

- Twitter IO archive: covers several countries up to a few months ago. Good for getting a sense of the size and 'feel' of typical nationstate twtter posts/ networks etc. https://transparency.twitter.com/en/information-operations.html
- Facebook ad library: contains all active ads that a page is running on Facebook products https://www.facebook.com/ads/library/ (About the Ad Library)

1.3 Collecting your own data using tools

The datastreams above will help you get a sense of what's known about the artefact and/or theme that you're investigating, and sometimes that's enough to craft a response (e.g. if there's a WHO page on a known scam, that might be enough evidence to ask for takedowns etc). But most of the time, you'll have to go collect your own data from across social media, and sometimes beyond (e.g. for paper flyers, we asked people if they'd seen them in their neighbourhoods too).

Where you collect from, and what you collect will depend some on the artefacts you found, but here are some of the ways.

1.3.1 Twitter data

Twitter data is studied a *lot* precisely because it has a lovely API. Since we use a lot of Python here, let's talk about Python libraries. If you have twitter API codes, then Tweepy is a good choice. If you don't want to use the twitter API, try Twint.

Various researchers post twitter data-gathering tools online. Andy Patel's twitter-gather is good if you're doing twitter network analysis https://github.com/r0zetta/twitter_gather

We have code based on an early version of Andy Patel's twitter_gather code in the github repo. It's andy_patel.py - call it with "python andy_patel.py name1 name2 name3 etc" where name1 etc are the hashtags, usernames, phrases (phrases in quotes) that you want to search Twitter for. Andypatel.py creates a set of files in directory data/twitter/yyyymmddhhmmss_hashtag1 etc with the tweets, most prolific urls, authors, influencers, mentions etc and gephi input data so you can create user-user etc graphs (see the gephi instructions in this BigBook for how to do that).

Data for earlier investigations are in the repo folder data/twitter if you want to see what that looks like.

1.3.2 Facebook data

The Facebook API is horrible. Most everyone tracking social media uses a third party like CrowdTangle (which isn't free) or scrapes for the data they want.

1.3.3 Reddit

Reddit data is regularly dumped in an easy to read format. For quick-looks, there are tools like https://www.redective.com/

1.3.4 Multi-platform tools

Reaper collects from a set of social media feeds. Trying that out.

Access tokens:

- Facebook: look at list in https://developers.facebook.com/docs/facebook-login/access-tokens/ - then used https://developers.facebook.com/tools/explorer/ to check token worked before putting into reaper.
    - "Page Public Metadata Access requires either app secret proof or an app token" - see https://developers.facebook.com/docs/apps/review/feature#reference-PAGES_ACCESS
-