12. Open Source Intelligence

TL;DR OSINT

- Image: reverse image search; scrape text, urls, hashtags, phone numbers, icons etc and search for each of these
- Tweet: download and analyse twitter search, check images, urls, main accounts from it. Check other social media for related accounts, groups, text
- URL: check registration with icann/whois, use builtwith to look for owner and related sites, use crowdtangle to look for social media mentions. Check site for social media links, owners, sales
- Facebook group: log group name, look for connected groups, look for URLs and other social media accounts, search internet and other social media sites for group name and/or group description text, check page owner / admin

Introduction

We use OSINT to investigate artifacts. Artifacts are the things we can see online. Common artifacts include:

- Tweets
- Twitter accounts
- Facebook groups

- Domains (websites)
- Hashtags
- Images
- Videos
- Audio fragments (e.g. voice messages)

Tracking artifacts helps to understand what's happening in an incident, how everything in it fits together, and what we can usefully pass on as information about it at the incident level, or usefully do to influence it.

Artifacts also include combinations of other artifacts. These include

- Domain networks
- Account networks (including botnets)
- Narratives

Many incidents start with an artifact that we investigate, finding connected artifacts, incident objects (e.g. actors and narratives) and techniques in the data available online in social media and other open sources (e.g. typical OSINT inputs). This chapter looks at some of the things we do when we meet different types of artifact.

General Advice

When you start investigating an artifact, check if anyone else is already tracking it. Check places like reddit: you might save yourself time.

The basic questions: What is this thing. How is it impacting the things we care about? Are there other teams doing something about it? What can we do about it? How much impact can we make in the things we care about, for the resources we need to expend?

Time can be confusing. Use ISO8601 format for dates where possible: yyyy-dd-mm, and either use UTC, or state the timezone you're using.

Handling Domains (URLs)

A url is a web address, e.g. https://www.washingtonpost.com/policies-and-standards/. A domain name is a link to a website, e.g. https://www.washingtonpost.com/. Within the domain name is a "top-level domain", e.g. washingtonpost.com. Domain names and URLs are useful things to track: most financially-motivated disinformation needs a URL to make money, and URLs are consistent: unlike hashtag spellings, they don't usually have variant spellings etc.

So you've got a URL. Now what? Well, you probably want to know about the URL - who created it, when, what's it connected to etc.

Google Dorks

Google Dorks are web searches that use Google's advanced search options. Using Google Dorks to Check Primary sources (from Henk van Ess's Finding patient zero):

"Step 1: Look at the link

- *Ex. https://www.sec.gov/litigation/apdocuments/3-17405-event-11.pdf*
- *Pull out just the domain name and Top Level Domain (Ex. sec.gov)*

Step 2: Use "site:"

- *Go to a generic search engine.*
- *Start with the query ("Dutch police") and end with "site:" followed directly with the URL (no spaces).*

- *Ex. "Dutch police" site:sec.gov*

Step 3: Adapt the "primary source formula" to your needs

- *Include specific folders (Ex. "Dutch police" site:sec.gov/public)*
- *Predict folders you think might be there*

Following the trail of Documents

Step 1: Establish the document type

- *Is it a doc \| pdf \| xls \| txt \| ps \| rtf \| odt \| sxw \| psw \| ppt \| pps \| xml file?*
- *Use filetype: and the type of file with no spaces (Ex. "filetype:pdf")*

Step 2: Include a phrase you'd like to search with in the document (could include a date)

- *Ex. You're searching for an invitation to an event from May 13, 2014, event. (Be sure to search for both the cardinal and ordinal forms, May 13 and May 13th.)*

Step 3: Who is involved?

- *Do you know the creator/host and it's website?*
- *Ex. The organizer is "Friends of Science" and its website is friendsofscience.org.*

When you combine all three steps, the query in Google will be:

"May 13th, 2014" filetype:pdf site:friendsofscience.org

Filtering social media for primary sources

Process for investigating the authenticity of a website :

Web searching a domain: Since we want to find out what other sites are saying about the site while excluding what the site says about itself, we use a special search syntax that excludes pages from the target site

- *Search syntax is website -site:website*
- *(Ex. baltimoregazette.com -site:baltimoregazette.com)*
- *Scan the set of results looking for sites we trust"*

Finding site ownership and connections

Enter the domain name into WHOIS Domain Tools or https://lookup.icann.org/lookup. Note who the domain was registered to: unfortunately, WHOIS blockers have dramatically reduced the value of WHOIS searches, so you may only find a proxy. Note when the domain was registered.

Use a backlink checker like ahrefs or smallseotools to see all the websites that link to the domain. Check the domain name on builtwith.com. If you're lucky that will tell you when and who. It will also tell you which sites have the same tags as this site: this helps you find connected sites. CogSecCollab code run_builtwith.ipynb produces the same results, but gives you json and a dataframe of those connected sites.

Site contents

Are there phrases you can use in a googlesearch, to find related objects? Run the search that allows repeated results, to see identical pages. About and terms pages are usually good places to look for these. Are there people or companies connected to the site? Start searching for them. CogSecCollab code googlesearch_for_terms.ipynb searches for terms/ pages.

Related Websites and Typosquatting

Astroturfers try to cover an area, whether it's geographical or demographic, and if they're doing it for money, they'll usually have multiple sites. Look at the title and url of the site. Do they have elements that might be repeated? Think about geography, verticals, and other clues there might be variants of this site. E.g. if you have xxxmichigan.com, check for the same pattern with other states' names, e.g. xxxwisconsin.com.

Typosquatting is when you create a site whose url is almost the same as a real or well-known one, often using combinations of letters (e.g. 'nn' instead of 'm') or urls (e.g. .gov.us) to fool people on a casual glance. Useful python libraries for finding typosquats include dnstwist for generating typosquats, and SnaPy for finding near-duplicates.

Looking for search terms in new domainnames can help spot new trends. whois newly-registered-domains is a list of domains created each day. Github code check_new_registrations.ipynb searches for strings of interest in that domain list. Newly registered alone isn't really an indication of anything; domains that are newly registered and active all within 24hrs, are worth watching, as are recently active and questionable domains. We have e.g. the Zetalytics API for searching through those.

References to the domain

Check social media - are there references to the URL, or groups / pages / accounts with the same name? Crowdtangle's chrome extension will give you a list of references to a site you're looking at, on Facebook, Twitter, Instagram and Reddit.

If there are references to the URL, are there common hashtags, phrases or people in common you can use to search for more sites?

Examples of tracking URLs include the references in [Data Safari rough notes: "pink slime" network].

Handling Tweets

Hashtags

"what do we consider worthy of collecting from twitter?" - FrankC

Good question. The TL;DR is that the reason we use the code that we do (andypatel_gettwitter.py from CSC tracking repo) is because we're looking for the objects that dominate and are related to the hashtag:

- we want to know which users are promoting it
- Which other hashtags are used heavily with it
- Which users on the hashtag are in suspicious configurations - e.g. one user linked out to lots of other people who aren't connected to each other (that's someone either pushing or pulling, depending on the direction of the links), or groups of users connected heavily to each other but not to anyone else on that hashtag (typical configuration for a botnet)
- we want to know which URLs are associated with the hashtag - if this is being used to make money, that money has to come from somewhere, and that's usually either online advertising, merchandise or paid services: either way, each of those is going to have a web address associated with it, and any grifter worth their salt is going to be pushing that address heavily

- We also collect images - that gives a good idea of what the themes are, because most good disinformation merchants know that images are more often exchanged than text. That's why you see all those posters with text on

The finding the configurations part - we use Gephi to look at the network; botnets and distributors stand out like little flowers in a Gephi network. But we could use networkx to do the same thing. There are also a set of tools in OSOME that will help you examine relationships quickly.

Raw data is useful too - it's where we start. But really, in social engineering, it's the relationships that count.

Botnets

I use bot sentinel and tools like it - ones like Hamilton68 monitor accounts from nation state actors (Russia, China etc - think embassy twitter feeds, RussiaToday etc), ones like Botsentinel monitor accounts active in earlier campaigns that might or might not be bots. The most valuable thing they give you is trends: what the recent chatter online is.

Bot detection is an art now. Once upon a time, it was as easy as "there are 100 accounts posting all the time, and they're all posting the same text", and finding them was basically "look for the Qanon hashtags". Now it's more subtle. There are some rules of thumb, like being suspicious of anything tweeting more than 100 times a day, but there's more to it, and a bunch of tools to help.

Chasing an image

There are a few things you're going to want to do with an image:

- Extract the text from it

- See where else it exists online
- Check to see if it's been altered / is fake

Extracting text: You can usually extract text from images using optical character recognition, OCR. There are libraries like Tesseract that can be called from Python (as e.g. pytesseract), but they have mixed results. A more reliable way to do this is to use the OCR built into search engines to pull the text from each image: yandex.com appears to be best at this (although always check because OCR still doesn't produce perfect results) but is Russian: if that's an issue for you, bing.com image search does this too.

Seeing where else an image is online:

- Mostly you'll be doing this by hand for new images, but a good first check is to see if an image (e.g. a photo) has been reused from an earlier event. Reverse image search from yandex.com and bing.com works well - tineye.com will call all the big image search engines for you (and you can laugh at some of the things they return...).

Checking for alterations: Bellingcat are the masters of online image forensics, and have a good guide to this Bellingcat guide. Look at tools like FotoForensics.

Video and Audio

Video

InVID_EU

"YouTube's search tool has a problem: it won't let you filter for videos that are older than one year. To solve this,

- *In a Google search include the keywords and site:youtube.com*
- *manually enter the preferred date into a Google.com search by using the "Tools" menu on the far right*
- *Then select "Any time" and "Custom Range." "* - finding patient zero

Audio

You can save an audio file from Facebook Messenger. The workaround is to use m.facebook.com on a Chrome browser - NOT on mobile. Click on the messenger icon. Go to the chat that has the audio. Right mouseclick on the (...) at the end of the message and you'll have the option to "Save Audio As".

Handling Facebook Groups

A lot of Covid19 disinformation is happening and/or moving at some point through facebook groups. We've been tracking some of these by hands whilst working out how to automate creating watchlists of groups, pages, accounts to check for new disinformation incidents forming before they hit the mainstream press.

Some academic references on this, focussed on antivax (one of the best-known and well-studied modern conspiracy theories).

Further Reading

Tracking facebook groups

- The online competition between pro- and anti-vaccination views
- Hidden resilience and adaptive dynamics of the global online hate ecology
- "New online ecology of adversarial aggregates: ISIS and beyond" with supplementary materials