Chapter 2 Disinformation Models

TL;DR Disinformation Models

We're using adapted information security standards to describe disinformation incidents, so we can share them with a large number of responders.

An incident is a coordinated sets of activities, over a relatively-short timespan, usually with an individual or team behind it. We describe these in terms of narratives (the storylines in the incident), TTPs (techniques used), and incident objects (actors, tools etc). We use STIX to describe most incident objects, AMITT to describe techniques, and text to describe narratives.

Deconstructing Disinformation

Part of CogSecCollab's work has been to model and populate the processes, tools, and countermeasures needed to respond in a distributed collaborative way to disinformation at scale. This chapter looks at some of the underlying models we use to understand and describe disinformation.

Disinformation Models

There are overlaps between disinformation, information security, machine learning, and military competition short of armed conflict. Whilst ideas from all of these appear feature in this book, the models and processes that we base most of our work on come from information security.

The big ideas here are:

- information security and disinformation defence are so similar that we can use the same tools for them both.
- If we have a common description language, we can share information about disinformation incidents in real time
- If we describe the moves disinformation creators use, we can mitigate or block those moves

This chapter covers mapping misinformation attack and defense into existing infosec frameworks, and how that mapping allows us to plan misinformation defenses and counters, assess tools and mechanisms, and handle the types of large-scale adaptive threats that machine learning applied to information security, MLSec, makes possible. This should help misinformation response move past the intelligence and passive defense stages of the SANS sliding scale, to architectural and more active defenses.

Cognitive Security

Information Security has always had three main layers (cognitive security, physical security, cybersecurity), but the cognitive one has been downplayed for a long time. For example, if you look at definitions of information warfare, they include electronic warfare, military deception, operation security, computer network operations, and psyops, where CNO is what we typically think of as infosec, and psyops as including disinformation.

Cognitive security is a rapidly growing domain that interacts with cyber and physical security, and includes things like information operations and disinformation. This covers tools, techniques and resources for threat sharing and response, and practical applications.

We do have a name clash here, but not a severe one: some groups use "cognitive security" to mean using AI in information security defence, but that's not often heard in MLsec communities.

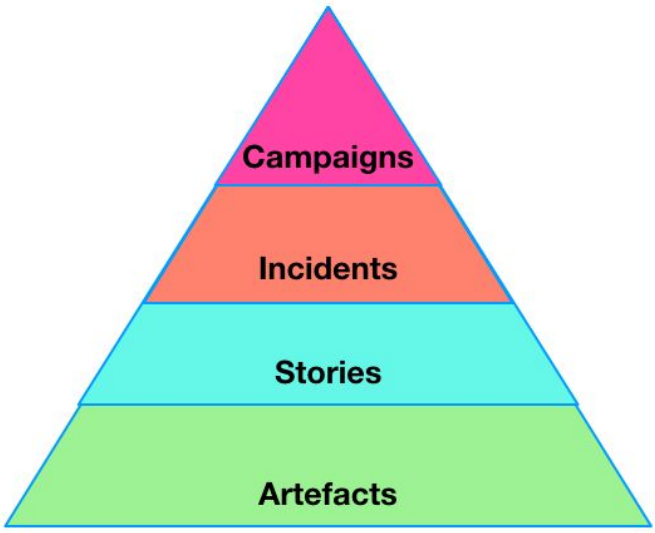
In here, when we talk about cognitive security, we're talking about one of the three main layers of information security:

- cybersecurity (the machines and networks infosec you usually read about),
- physical security (physical, as in breaking into the building because it's sometimes easier to steal from the computer than break in electronically) and
- cognitive security (attacking and defending human minds and the networks between them as part of an infosec attack).

Disinformation is an attack in the cognitive security domain, but there are others that can be used - social engineering is getting humans who are part of an information security system to help you break that security (e.g. by letting you into their systems). The work that misinfosec and others did on how attacking human beliefs and emotions can be viewed and defended in similar ways to attacks on machines is, broadly speaking, Cognitive Security.

Disinformation Layers

Disinformation Pyramid



Disinformation Pyramid

We use the disinformation pyramid to show how we connect information operations, threat intelligence, osint research and disinformation data science.

- Campaigns: are long-term disinformation operations. They're focussed around a theme, like specific geopolitics (e.g. "make everyone like china" or "Ukraine is really Russia"), and are often nation-state-funded, but might also be from interest groups (e.g. far-right-wing, antivaxxers etc). Information operations work is often at this level.
- Incidents: these are the short term, cyclic things we track. They're coordinated sets of activities that happen over a defined timespan that usually indicates some form of team or individuals driving them. Incidents have things with defined parameters like TTPs that we can share, threat actors, and other objects that you'd recognise from TI, but also including context and narratives. OSINT research and threat intelligence usually happens on this level.
- Narratives: are the stories that we tell about ourselves and the world. They're stories about who we are, who we do and don't belong to, what's happening, what's true (e.g. Covid19 was caused by 5G masts). Tagging information with defined narratives make it easier for us as analysts to follow the flow of information across the internet and beyond.
- Artifacts: Incidents and Narratives show up online as artefacts: the text, images, videos, user accounts, groups, websites etc and links between them all that we collect and use to understand what's happening. Data scientists usually start here.

So what looks to outside observers like analysts simply hunting down a hashtag or a URL, describing a narrative, or trying to understand the things that link to it is so much more; it's really a part of creating an inventory of the discrete elements of each incident, or the objects used by a disinformation team or campaign, so we can a) share a summary of what we think is happening, and b) disrupt both those component parts, the TTPs behind them, and the incidents and campaigns they support.

Actors in the Pyramid

For power-motivated disinformation, we have three main groups of people: the creators of misinformation ('attackers'), the people trying to counter them ('defenders'), and the targets of the misinformation ('populations'). Typically, attackers start at the top of the pyramid and work their way down. Defenders are at the bottom and work their way up.

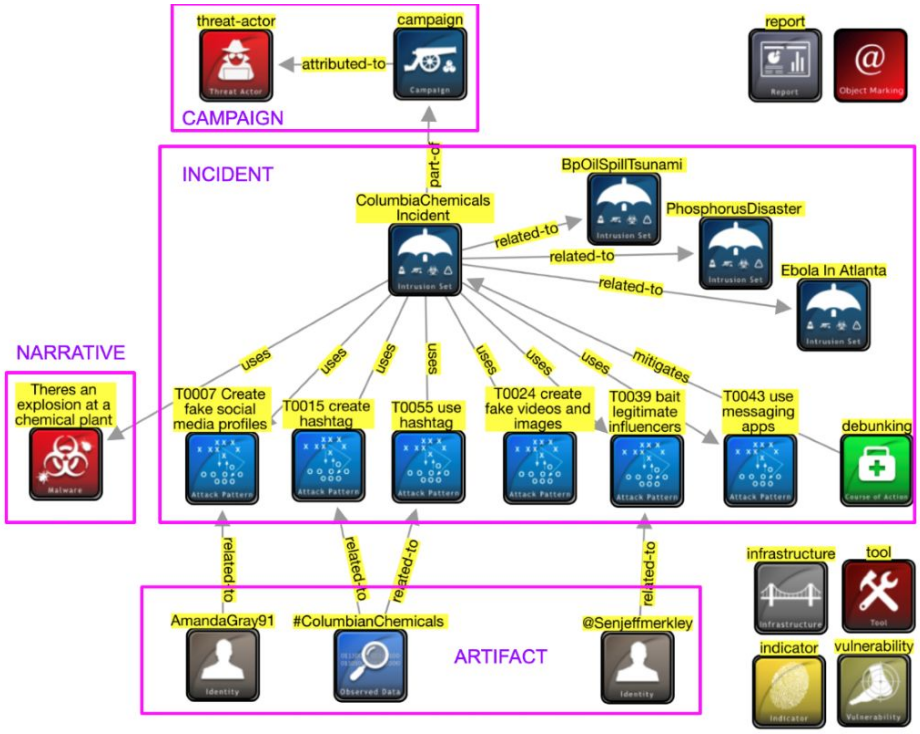
- Red. Attackers create incidents (e.g. Macrongate), which often form part of longer-term campaigns (e.g. destabilize French politics). Human communication is generally at the level of stories, or narration: we tell each other stories about the world, as gists or memes. And to tell these stories, we need artifacts: the users, tweets, images etc that are visible in each attack. Attackers have a goal they want to accomplish and design a misinformation campaign to achieve that goal. They manufacture one or more incidents, each incident has its own narrative which is told through a series of artifacts. Those artifacts can be posts, tweets, stories, deep fakes, etc. As attackers move down the pyramid, more work must be done. A single campaign can have thousands of artifacts transmitted by tens-of-thousands of accounts.

- Blue. Whilst the attacker sees the whole of the pyramid from the top down, the defender usually sees it from the bottom up, working back from artifacts to understand incidents and campaigns, unless they're lucky enough to have good insider information or intelligence. Most current misinformation work is at the artifact level, although there has been narrative (story) level work happening recently. By contrast, defenders start at the bottom of the pyramid. They see an artifact, and then another, and at some point, they may be able to tie all of these artifacts into a cohesive narrative. Eventually several of these narratives can be tied to distinct incidents and with enough investigation and perhaps a little attribution, a campaign can be discovered. This is definitely an "uphill" climb. Defenders will never uncover ever artifact and are likely to miss numerous narratives and incidents because they simply don't have access to the communities and platforms where they present. Even with access, they may never get around to analyzing the information or even recognize it as linked to a campaign.
- Non-team. This is cognitive security, so there are many other actors in the pyramid, including people unwittingly sharing disinformation, or being the targets of disinformation narratives.

When you look at that pyramid, those layers aren't just about information - they're also about action, and understanding how to tie together both attack and defence activities from different layers. Information operations tends to be at the top layers, data science at the bottom: people doing this work need to be able to talk to and work with each other.

Disinformation Objects

STIX and extensions



STIX diagram for Columbia Chemicals

STIX is a data standard used to share information between threat intelligence organisations like ISACs. It's a rich language that describes threat objects and the relationships between them, is extensible, used by existing threat intelligence sharing communities (ISACs, ISAOs etc) so we'd be patching into an existing sharing system. It's also supported by and integrates well with existing community-supported, open-source tools.

Misinformation STIX	Description	Level	Infosec STIX
Report	communication to other responders	Communication	Report
Campaign	Longer attacks (Russia's interference in the 2016 US elections is a "campaign")	Strategy	Campaign
Incident	**Shorter-duration attacks, often part of a campaign**	**Strategy**	**Intrusion Set**
Course of Action	Response	Strategy	Course of Action
Identity	Actor (individual, group, organisation etc): creator, responder, target, useful idiot etc.	Strategy	Identity
Threat actor	Incident creator	Strategy	Threat Actor
Attack pattern	Technique used in incident (see framework for examples)	TTP	Attack pattern
Narrative	**Malicious narrative (story, meme)**	**TTP**	**Malware**
Tool	bot software, APIs, marketing tools	TTP	Tool
Observed Data	artefacts like messages, user accounts, etc	Artefact	Observed Data
Indicator	posting rates, follow rates etc	Artefact	Indicator
Vulnerability	Cognitive biases, community structural weakness etc	Vulnerability	Vulnerability

Disinformation version of STIX

STIX translates well for disinformation use. We added two objects to STIX for disinformation: incident, and narrative, and didn't need to change anything else. We use custom objects to represent these fields and be OpenCTI compliant.

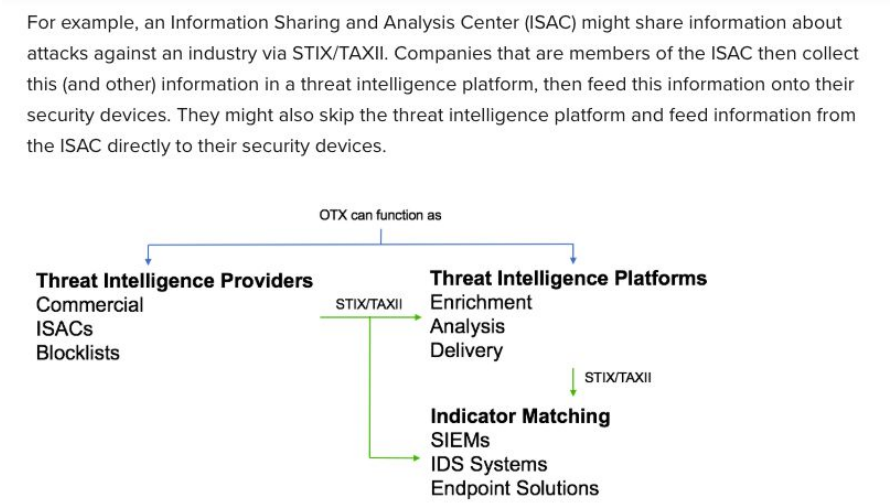


Image from https://stixproject.github.io/about/STIX_Whitepaper_v1.1.pdf

AMITT is now available as a STIX 2.0 bundle, from https://github.com/cogsec-collaborative/amitt_cti. When STIX 2.1 delivers an incident object we'll migrate to that.

References

STIX

- https://www.alienvault.com/blogs/security-essentials/otx-is-now-a-free-stix-taxii-server
- https://pukhraj.me/2019/01/27/what-does-a-national-cyber-shield-look-like/#more-861
- Einstein program
- STIX… and OpenC2
- https://stixproject.github.io/about/STIX_Whitepaper_v1.1.pdf
- https://threatconnect.com/stix-taxii/
  https://www.crowdstrike.com/blog/indicators-attack-vs-indicators-compromise/
- https://oasis-open.github.io/cti-documentation/stix/intro?_ga=2.135668339.378020639.1559740731-781460544.1559740731

Disinformation Typographies

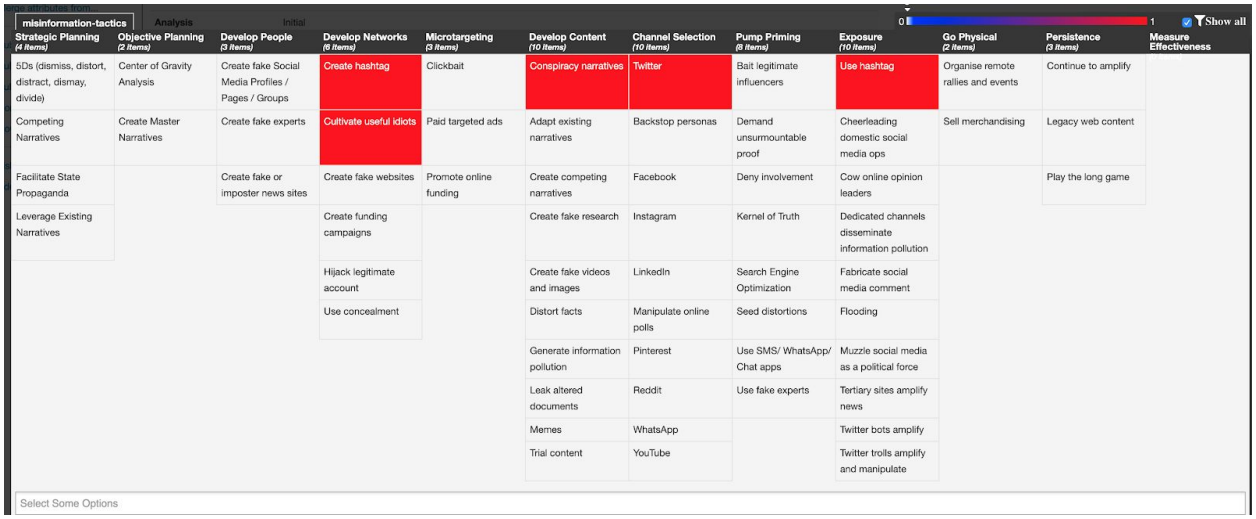
STIX gives us objects, e.g. threat actor, but doesn't give a standardised way to describe the type of each actor, e.g. nationstate threat, for-profit threat, etc. We're working on that, with NATO, based on DFRlab's Dichotomies of Disinformation.

References

- https://github.com/DFRLab/Dichotomies-of-Disinformation#appendix-a-codebook

Disinformation TTPs: Tactics, Techniques, Procedures

AMITT Framework



AMITT TTP Framework, as seen in MISP

One of the disinformation objects that gives us a lot of information is the TTPs (techniques, tactics, procedures). In 2019, the Credibility Coalition MisinfosecWG team built a disinformation equivalent to the ATT&CK framework:the AM!TT (Adversarial Misinformation and Influence Tactics and Techniques) TTP framework, incorporating components from existing infosec standards, misinformation models, psyops, and marketing models (e.g. sales funnels), and designed using a wide range of example incidents, ranging from nationstate to small-group in-country operations. AM!TT's language and style is adopted from the MITRE ATT&CK framework, and its form is designed so we can use all the tools available for ATT&CK on it. CogSecCollab continues to be involved in the evolution and maintenance of AM!TT, including the use of subtechniques in the model.

AMITT is designed to give responders better ways to rapidly describe, understand, communicate, and counter misinformation-based incidents. We use the AMITT framework to break each disinformation incident down into its component TTPs, and to design and use

TTP-level countermeasures. It's designed as far as possible to fit existing infosec practices and tools, giving responders the ability to transfer other information security principles to the misinformation sphere, and to plan defenses and countermoves.

The latest version of AMITT is held in the Github repository https://github.com/misinfosecproject/amitt - in there, you can view a populated framework, where you can click on a technique and get details about what it is, who uses it, and which counters are available for it.

Every AMITT component has a unique id (e.g. T0018 Paid targeted ads). The framework is read left-to-right in time, with the entities to the left typically (but not necessarily) happening earlier in an incident. Its components include:

- Phases (not shown): higher-level groupings of tactic stages, created so we could check we didn't miss anything. The phases are separated into left-of-boom (planning, preparation) and right-of-boom (execution, evaluation), to represent activities before (left) and after (right) an incident is visible to the general public. The tactics below each phase belong to that phase.
- Tactics (top row): stages that someone running a misinformation incident are likely to use
- Techniques (all other rows): activities that an incident creator might use at each stage. The techniques below each tactic belong to that tactic. An example of a technique is T0010: Cultivate ignorant agents. This describes pulling in unwilling agents - through hiring them, or co-opting through emotion, agenda, sympathy (eg. conspiracy theorists are often ignorant agents). The technique doesn't define how to achieve this. There are many ways to hire or co-opt individuals, each potentially requiring its own counter.
- Tasks (not shown): things that need to be done at each stage. Tasks are things you do, techniques are how you do them.

To use AMITT, list and share the components you see in your incident: AMITT is now built into the MISP tool, making this easy to do. Compiling and reporting incidents is an important aspect of both responding and developing the tools needed to do so. To be effective, those reports should include as much information as possible about the stages and techniques at play in those incidents.

MITRE ATT&CK 2.0 is in the works and it refactors high level capabilities into implementations. It's a direction we'd like to see with AM!TT and something our group will continue to work toward.

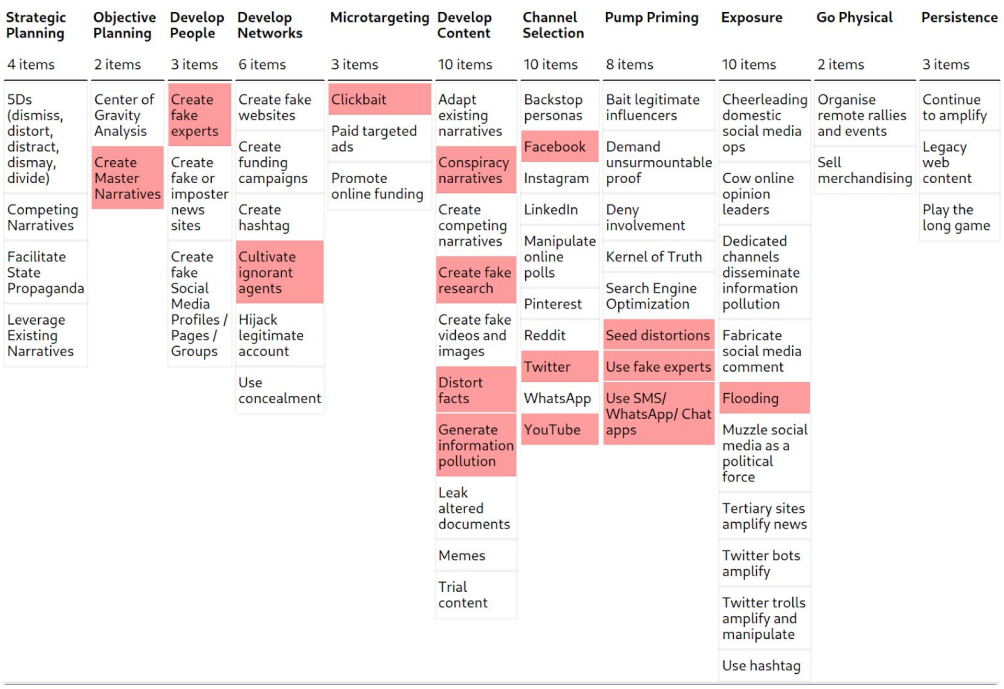
Example: Double Diamond

We saw the use of T0010 (Cultivate ignorant agents) in the Double Deceit incident. This technique was used by EBLA, registered as an NGO in Ghana, doing NGO charity work on the

ground in Ghana (stationary to students). Problem is that that it's actually a Russian front. They hired local youth to post on social media. These people were not aware that they were part of a Russian troll farm. They also attempted to co-opt legitimate US influencers, to retweet their messages. Content was not political, and this appeared to be audience building.

We believe this is associated with the IRA. It appears to have been a failure.

Example: Plandemic



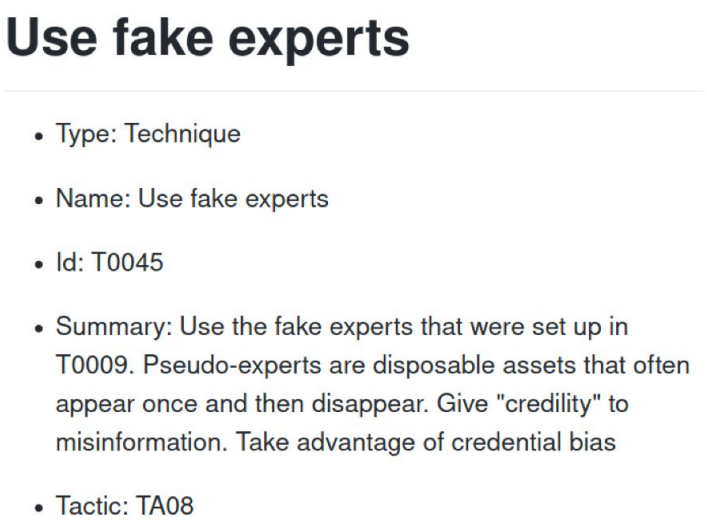
Strategic Planning	Objective Planning	Develop People	Develop Networks	Microtargeting	Develop Content	Channel Selection	Pump Priming	Exposure	Go Physical	Persistence
4 items	2 items	3 items	6 items	3 items	10 items	10 items	8 items	10 items	2 items	3 items
5Ds (dismiss, distort, distract, dismay, divide)	Center of Gravity Analysis	Create fake experts	Create fake websites	Clickbait	Adapt existing narratives	Backstop personas	Bait legitimate influencers	Cheerleading domestic social media ops	Organise remote rallies and events	Continue to amplify
Competing Narratives	Create Master Narratives	Create fake or imposter news sites	Create funding campaigns	Paid targeted ads	Conspiracy narratives	Facebook	Demand unsurmountable proof	Cow online opinion leaders	Sell merchandising	Legacy web content
Facilitate State Propaganda		Create fake Social Media Profiles / Pages / Groups	Create hashtag	Promote online funding	Create competing narratives	Instagram	Deny involvement	Dedicated channels disseminate information pollution		Play the long game
Leverage Existing Narratives			Cultivate ignorant agents		Create fake research	LinkedIn	Kernel of Truth	Fabricate social media comment		
			Hijack legitimate account		Create fake videos and images	Manipulate online polls	Search Engine Optimization	Flooding		
			Use concealment		Distort facts	Pinterest	Seed distortions	Muzzle social media as a political force		
					Generate information pollution	Reddit	Use fake experts	Tertiary sites amplify news		
					Leak altered documents	Twitter	Use SMS/ WhatsApp/ Chat apps	Twitter bots amplify		
					Memes	WhatsApp		Twitter trolls amplify and manipulate		
					Trial content	YouTube		Use hashtag		

Plandemic TTPs in AMITT

The AMITT framework was built to be practical. We need to be able to translate our findings into an actionable story.

Plandemic is a debunked conspiracy theory video which makes some false claims about the nature of COVID-19. Despite high production quality the self reported cost to produce the film was less than $2000. Zach Vorhies, an individual associated with QAnon, claims to be the social media marketer behind the viral success of the video. NYT reported his GoFundMe campaign titled "Help me amplify Pharma Whistleblower Judy Mikovits."

We can map out this small, but successful, operation in the AMITT framework to help us understand what capabilities the actor has and potentially how they're resourced. As with ATT&CK, we can start building an understanding of actors' capabilities over time.



Technique T0045, used in Plandemic

Plandemic exploited credential bias, and relied heavily on AMITT technique T0045: use fake experts. The "expert" here was Judy Mikovits. Some of the narratives used included that vaccines contain Covid19 virus, masks activate Covid19 virus, and that the Plandemic video was exposing scientific and political elites.

Fake experts are interesting because their credentials lend credibility to outrageous claims. Fake experts use their credential suspend disbelief. Fake experts create an illusion of “another side” of the argument (anti-vaxx, climate change, etc.). It’s an effective technique in part because it's a human story. It plays into the narrative of a lone researcher, an outsider, bravely facing off against the scientific and political elites who seek to destroy her and her reputation to maintain the status quo, and it’s a story that’s cast her as a victim.

Reading

AMITT

- Walker et al, Misinfosec: applying information security paradigms to misinformation campaigns, WWW’19 workshop
- http://overcognition.com/2019/05/13/misinformation-has-stages/
- https://medium.com/misinfosec/disinformation-as-a-security-problem-why-now-and-how-might-it-play-out-3f44ea6cda95

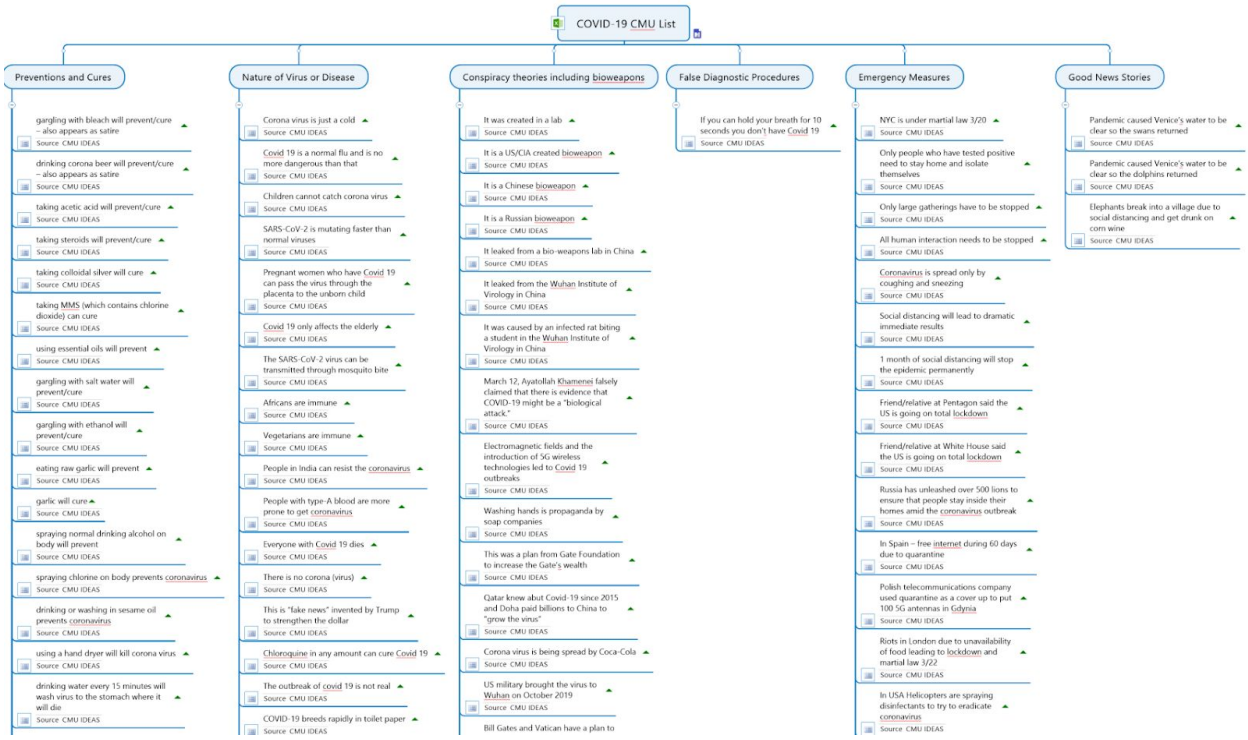
Techniques

- [Russian Election Trolling Becoming Subtler, Tougher To Detect]
- [Big Lies and Rotten Herrings: 17 Kremlin Disinformation Techniques You Need to Know Now]

Social Media Object Set

STIX gives us artifact object types Observed Data and Indicator, but in MISP we get into more detailed object types like email, url. MISP didn't have a set of objects to cover social media data, so we added a new set with a new object for each new platform type (twitter-post, facebook-group etc). We initially tried using generic objects (social-post, social-group etc), but found these confusing and difficult to work with at speed.

Narrative Models



Mindmap of Covid19 Narratives

We know we need to track narratives as they form, combine with other narratives, die away and sometimes reemerge, but we haven't settled yet on a good representation for this. We've tried

mindmaps, and looked at how to match known narratives with the results of things like text-based clustering and anomaly detection.