From Words to Intelligence: Leveraging the Cyber Operation Constraint Principle, Natural Language Understanding, and Association Rules for Cyber Threat Analysis

Ronan Mouchoux[1] and François Moerman[1]

[1]XRATOR

This paper was presented at Botconf 2023, Strasbourg, 11-14 April 2023, www.botconf.eu
This paper is published in the Journal on Cybercrime & Digital Investigations by CECyF, https://journal.cecyf.fr

Abstract

This paper proposes a system for collecting and structuring blog articles about cyber-attacks, with the goal of improving the ability of security researchers to compare threat actor modus operandi.

By grounding our work in the field of criminology, we also formulate a Cyber Operation Constraint Principle that could inform future research. We derived from it a tool, the AbductionReductor, that has the potential to augment partial knowledge about a threat actor's behaviour while investigating its actions.

Our approach has the potential to significantly support cyber threat analysis and investigation. Future research must focus on the challenge of synchrony and diachrony linguistic analysis.

Keywords: Criminology, Computational Cyber Threat Intelligence, Natural Language Processing, Modus Operandi.

1 Introduction

Cyber Threat Intelligence (CTI) creates operational knowledge about a situation that evolves because of technological or business evolution, the attacker landscape, or the defender posture [1].

The sub-discipline of Tactical CTI [2] – also referred as operational CTI [3] - focus on providing information about the adversary behaviour, during the pre-exploitation and post-exploitation phases. To assist the structuration of investigation and restitution, Tactical CTI relies on frameworks such as Lockheed Martin's Intrusion Kill Chain or MITRE ATT&CK® and structured language such as OASIS STIX [4].

Most of the work remains manual and based on the analyst's prior knowledge and interpretation of the frameworks and language. The analysts' production is in natural written language. As a result, Tactical cyber threat intelligence is an ad-hoc process with variable results and no quality standard. There is no common practice in mapping threat events and objects with structured language and expression, both within and across an organization. This implies a decrease in the operationalization of Tactical CTI analysts' production for its own future usage, for incident response and for attack detection [5].

This lack of standardization and structure in Tactical CTI poses a significant challenge to defenders, who must constantly adapt to new threats modus operandi, being forced to rely only on their own expertise, manual work, intuition, and long debate with peers to be able to match two different modus operandi narratives.

In recent years, there has been an increasing interest in using natural language processing (NLP) and machine learning (ML) techniques to automate the processing and analysis of CTI data, with the goal of improving the speed, accuracy, and reproducibility of cyber threat analysis [6]. However, this approach often requires huge volume of data as input, which is lacking in Tactical CTI [7].

In this paper, we propose a system that addresses this gap by collecting blog articles about cyber-attacks, normalizing them with a defined vocabulary, and storing them in a structured language. We believe that this system has the potential to significantly improve the ability of security researchers to compare threat actor modus operandi.

Furthermore, by grounding our work in the field of criminology, we formulate a Cyber Operation Constraint Principle that could inform future research in Cyber Threat Intelligence. We derived from it a tool, the AbductionReductor, that we believe has the potential to augment partial knowledge about a threat actor's behaviour while investigating its actions.

Looking back at our journey in developing the system as well as the strong advancement by academics in applying NLP to Cyber Threat Intelligence, we trust that the research community must now assess the challenges and benefits of synchrony and diachrony linguistic analysis to historical base of threat reports.

2 Background

In this paper, we focus on the application of Natural Language Understanding (NLU) in cyber threat intelligence based on Postulate 1.

Postulate 1: a cyber-attack is governed by the combo "Threat Agent - Threat Event - Target - Remediation".

To contextualize our postulate, we conduct a review of NLU context for infectious disease medical research and criminal reports, where similar combos are used to describe respectively "*Pathogen Agent – Disease – Patient – Cure*" and "*Criminal - Acting Out – Victim – Rehabilitation*". We then move to a review of NLU in cyber threat intelligence to showcase the current state of the art in the field.

2.1 Medical Research

Medical research articles are predominantly using the introduction, methods, results, and discussion (IMRaD) structure [8]. This means that the document can be segmented and each segment can go through dedicated algorithms to extract expected information.

In addition, the World Health Organization provides references guides and best practices for naming diseases and pathogens [9]. Same as for polyfunctional organic molecules that are used to create cures [10]. This means that the identification of entities is relatively predictable, as unknown entities are generated from a closed taxonomical space.

The structuration of medical research articles goes deep to the linguistic features of sentences, resulting in schematic structure [11]. This means that the phrasing is relatively predictable and relationship between entities is easily captured with heuristics.

Even if the abstract problem of infection disease is relatable with the one of cyber-attack, the way medical research is conducted and structured facilitates its automated processing in regards with other research or investigation areas with more framing, phrasing, and naming diversity.

2.2 Criminal Intelligence and Criminal Justice Reports

Most of criminal justice reports are unstructured and narrative, built upon an eyewitness, victim or suspect recollecting their version of the event [12]. A concept that Goodchild called "citizens as sensors" and applies it to crowdsource geography [13]. This method of collection results in noisy content and various type of writing mistakes [14].

Such reports are relying also on specialized vocabulary and glossary, providing offence categories and mapping to law codes. But the incident, or modus operandi, is generally described in a narrative form [15]. It implies that two events can describe the exact same and discriminatory modus operandi, but the narrative form prevents easy information retrieval to link the two cases.

The quest to match similarities in criminal offense modus operandi is an old sea snake with several benefits. Fosdick in 1915 wants to go beyond the Bertillon system (“*organic Indicator of Compromise (IoC)*”) to augment crime detection [16]. He also points that habitual criminals commit their crime using the same method over and over and do not switch to unfamiliar methods. Borrowing the “Script” concept from cognitive science, Cornish argues that the knowledge about the procedural aspects and procedural requirements of crime commission has the potential to enhance situational crime prevention [17].

An individual achieves the status of criminal by their action and not by their own essence. Crime reduction paradigms based on Crime Opportunity Theories, such as the Routine Activity theory [18], emphasize approach on early or proactive prevention and detection of criminal offense based on modus operandi data points, one of the most famous being the intelligence-led policing [19].

Text Mining and NLP activities are hindered by the secrecy, privacy and ethical natural barriers of police and justice activities. Notable works include entity extraction [20] or modus operandi topic modelling [14].

We have not been able to find research that tries to assess free-text modus operandi to a criminal offense terminology.

2.3 Cyber Threat Intelligence

The origin of Threat Intelligence seems to come from 1970s-1980s aircraft electronic countermeasures efforts [21]. The term Cyber Threat Intelligence, the concept as we know it today, seems to have been first coined in 2000 in a patent describing a system and a method for the collection, analysis, and distribution of cyber-threat alerts [22].

Cyber Threat Intelligence reports may take many forms such as a blog article, whitepaper, conferences talk, criminal report, short messages (“tweet”), or newspaper. The producers may be aficionados, private sector professionals, government agency or law enforcement.

Postulate 2: Cyber Threat Intelligence is rooted in Criminal Intelligence. As such the essential object of study in Tactical Cyber Threat Intelligence is the Threat Actor’s modus operandi, a.k.a. Tactic, Technique and Procedure (TTP).

We will review the literature to spot opportunities in the cyber domain that may not exist in the physical domain, but will take with a lot of caution the research results as the modus operandi matching problem seems to be still valid in physical crime and classical criminology.

2.3.1 From IoC to TTP

In physical space and traditional crime, Locard's exchange principle builds the first stone of modern forensic science with its concept of “Every contact leaves a trace” [23]. The same principle applies in the cyber domain [24].

The interpretation of forensics traces is an abductive and inductive process [25]. Only valid deductive inferences preserve truth, when strong inductive and strong abductive approaches lead to a conditional conclusion with some degree of confidence [26].

A software is a set of instructions that a human could perform manually. Software interactions creates digital traces. If an investigation is able to find a software involved in a cyber-attack, the abductive and inductive conclusion built from the forensic interpretations can be corroborated by deductive reasoning thanks to static and dynamic malware analysis.

A Threat Actor can either manually perform malicious instructions or use a malicious software that implements them. A set of sequenced malicious instructions, automated or manually executed, is defined as a Procedure. A procedure is a specific implementation of a Technique. A technique is a mean to achieve a tactical goal, a Tactic. [27]

In the cyber domain, you can only witness the action of the threat actor through its digital traces. Its behaviour, its intent, is a conditional conclusion with some degree of confidence brought by the investigator’s interpretation.

2.3.2 Extracting TTP

Pragmatical models such as the “Pyramid of Pain” by David Bianco or the “Detection Maturity Level

Model" by Ryan Stillions emphasize the effectiveness in the cyber domain of leveraging modus operandi analysis for threat actor deterrence and situational cybercrime detection. A principle that we agree with, in the light of our review presented in section 2.B. We then further understand the relevance of capitalizing and normalizing TTP description for investigation and defence purpose.

In 2016, ZHU et DUMITRAŞ proposes an automated approach for generating features to detect android malware, called FeatureSmith [28]. It uses natural language processing to mine security literature, identify behaviours associated with malware, and map them to testable features. The system achieves high accuracy and can suggest features that are not in manually engineered sets.

In 2017, Usari et al. focused on automating the extraction of Tactics, Techniques, and Procedures (TTPs) from unstructured sources to enable timely and cost-effective implementation of cyber defence [29]. The authors claim that their work, TTPDrill, is the first to address this issue with reasonable accuracy. They develop an automated analytics tool to extract structured TTPs from cyber threat intelligence reports, using a semi-automatic Threat-Action Ontology based on MITRE CAPEC and MITRE ATT&CK. The authors augment the tool with ActionMiner to further enhance its capabilities [30].

However, Ayoade et al., published after Husari et al.'s paper, is highly critical of the approach from TTPDrill, finding a 14.8% accuracy on the tactic's predictions, when using their datasets [31]. The paper aims to reduce search time for analysts who want to reproduce an attack type for defence evaluation. The authors construct an end-to-end system that includes collection and feature extraction, followed by classification of the report into ATT&CK tactics, and finally into kill chain phases based on rules. The authors claim that their approach outperforms TTPDrill, with up to a 78% increase in classifier accuracy. The authors choose to use a text classification approach rather than an ontology-based information retrieval method.

In 2020, Thein et al. proposed a method for extracting event information from security reports and estimating the kill chain phases in a paragraph-based analysis. The proposed model is trained with ATT&CK in experiments and achieved an average F1-score of 0.67 and an average accuracy of 65% in estimating the cyber kill chain phase, with an 86% recall for extracting core features [32].

In 2022, Lin et Hsiao proposed a system called PELAT that fine-tunes BERT model with 1,417 articles from the MITRE ATT&CK framework to enhance its attack knowledge [33]. PELAT transfers its knowledge to perform semi-supervised learning for unlabelled attacks network packets to generate their tactic labels, and it predicts tactics for new attack packets by processing their payload with a downstream classifier. The authors claim that PELAT can effectively reduce the burden of manually labelling big datasets and can achieve high precision, recall, and F1 scores on testing datasets, as well as identify over 99% of tactics on two other testing datasets.

Those methods focused on mapping TTP to MITRE ATT&CK tactics or MITRE CAPEC.

2.3.3 Mapping Threat Event to MITRE ATT&CK techniques

The ability to identify in plain text the narration of the TTP is a first step, but the ability to compare and match them is of greater value, as described in section 2.B. The challenge is to find the right dosage between abstraction and precision. Too much precision may reduce the ability to match similar modus operandi. Too much abstraction will make all modus operandi look the same and hinder the ability to discriminate them and attribute them to an operational entity.

Postulate 3: The right dosage between abstraction and precision to normalize TTP is at the MITRE ATT&CK technique level.

In 2019, Legoy published her thesis which discusses the development of a tool that can automatically extract ATT&CK tactics and techniques from cyber threat reports, inspired by Ayode et al. [34]. The tool achieved a macro-averaged F0.5 score of 80% for the prediction of tactics and over 27.5% for the prediction of techniques. She highlights the limited amount of labelled data as a major issue, and that the quality of the included reports may not be sufficient to predict all labels accurately. Additionally, the ATT&CK framework is bound to change, with the introduction of three levels: tactics, techniques, and sub-techniques. While tactics will stay the same, more general techniques will be created, and more precise sub-techniques will be associated with them. In 2020, she further developed her idea and published the open-source tool rcATT on github [35].

In 2021, MITRE Corporation released the Threat Report ATT&CK Mapper (TRAM) based on the procedure's description available for each MITRE ATT&CK techniques and logistic regression [36]. But the results map an entire paragraph with techniques proposal that have to be manually selected by a human.

In 2022, Alves et al. presented the results of using 11 different BERT models to classify sentences from MITRE's labelled sentences base into various techniques and sub-techniques of the MITRE ATT&CK framework [37]. The best performing model achieved an accuracy of 82.64% on the test dataset with RoBERTa Large model and 78.75% on the inference dataset with BERT Large Cased. The paper highlights the challenges associated with sentence-level classification due to the impreciseness of the original sentence and the presence of multiple techniques or sub-techniques in a single sentence. The authors manually reviewed some misclassifications and discussed mixing techniques and sub-techniques. Finally, the authors suggest that MITRE's database organization allows for a multiclass modelling, but the data itself also allows for a multilabel approach, as the same description of a technique may fall under several MITRE ATT&CK techniques.

2.3.4 Our contribution

We must stress that this paper is the result of an experiment for fun of two engineers, with a not very formal approach, with no quality metrics and no computing performance optimization. Our goal was to create a proof of concept of an end-to-end system that takes a web article as an input and provides a STIX formatted output, normalized with MITRE ATT&CK techniques, with relationship between STIX entities, with the fewest possible errors. We engaged in this project development based on an intuition that we have been able to formulate as a principle thanks to the criminology literature review and our expertise: the Cyber Operation Constraint Principle. We have turned our intuition into a tool: The AbductionReductor. We use for our project a single dedicated server (Intel W3520 - 32GB DDR3 ECC 1333 MHz). On this quest, our contribution is multiple, and, we hope, will inspire curious researchers.

We provide a literature review linking modus operandi analysis challenges and relevance in the cyber domain and in classical criminology / criminal investigation. It leads us to formulate the Cyber Operation Constraint Principle.

We conducted the test on a large dataset, composed of 17153 articles, gathered from twenty-six sources, covering the 2007-10-27 / 2020-09-22 period. In addition, we leverage MITRE ATT&CKv6.3 (Enterprise, Mobile and former PRE-ATT&CK), eight malware-oriented MISP Galaxies Clusters, one adversary-oriented MISP Galaxies Clusters, thirteen long texts for domain specific language disambiguation and a geo database for place name matching and plotting.

Instead of using a machine learning-based approach, our pattern matching methodology enhances reproducibility and explainability.

Our system collects and transforms into the relevant STIX2.1 Domain Object (SDO) CVE entity, MITRE ATT&CK techniques, victim and offender countries, threat actor names and aliases, malware names and aliases, report.

Our system automatically aggregates the articles talking about the same malware or threat actor but mentioned across several aliases.

Our system provides the ability to automatically exclude threat actor names and malware names that are too generic and may provide false positives.

We use Dependency Parsing (Subject-Verb-Object, SVO) to explore the grammatical structure of a phrase to extract the relationship between entities (e.g.: a given threat actor using a given malware) and classify them following the STIX 2.1 Relationship Object (SRO).

Our result is an end-to-end system that takes as input an article HTML and produces a STIX2.1 formatted JSON.

We offer to reduce the abductive effects on investigation by augmenting techniques identified in a report with probable techniques using the A Priori association rule algorithm. This also applies for details about the attacker that are kept secret by the author for competitive advantage or secret classification.

We offer a quantitative analysis on the mention of IOC, CVE, techniques, country, threat actor name and malware name extracted from the dataset.

We trust that our knowledge in cyber adversary analysis and cyber-attack simulation allowed us to produce a street smart approach that is only asking to be optimized by seasoned architects and developers and contribute to the nascent field of computational Cyber Threat Intelligence.

Researchers could gather insight to achieve quick wins and gain industrial adherence in their work, leading them to formalize tools or approaches relevant for the cyber security and the computational criminology practitioners.

3 Methodology

To design our system, we identify three main components:

- **Data Collector**: Gathers the articles and pre-processes them.
- **Entity Matcher**: Spots selected CTI relevant entities.
- **Analytical Portfolio**: The Analytics Portfolio that provides specific analytical tasks. Here we present the AbductionReductor, Entity Trainer, the QuantitativeAnalyzer and the STIXCrafter.

Those components are composed of modules, described below. We code in python. Our database are JSON files manipulated with the Panda library.

Learning the lessons from Medical Report NLP, we will parse each source (that has the same document structure) separately and will leverage naming convention to guide our entity recognition.

3.1 Source selection

Our selection was based mostly on the kind of sources we would happily read. We sometimes selected only a portion of the source feed, such as tag or categories, to closer fit our CTI perspective.

BlueLiv	PaloAlto - Unit42
CheckPoint	RecordedFuture
ClearSky	RiskIQ
Crowdstrike	RiskyBiz
DataBreachToday	Secureworks
WeLiveSecurity	Talos
FireEye	Telsy
Fortinet	ThreatPost
IBM X-Force	TrendMicro
Intel471	UK-NCSC
Securelist	US Dept of Justice
McAfee	Volexity
MeltX0r	Wired

Table 1 – List of sources

3.2 Data Collector

The Data Collector is responsible of identifying new blog articles, collecting their HTML form, extracting the article's text from the page's text and parsing the article text for meta data enrichment. It is an inline process for each source, to precisely fit each of its specificities:

- **Blog Indexer**: Crawling the blog "landing page" and collecting all article URLs listed, through the whole pagination. It creates an index of all available articles. As each blog may have its own pagination scrolling technique, we have one indexer per source.
- **Article Crawler**: Collecting the raw HTML of a page.
- **Article Cleaner**: Automated and generic boilerplate removers are not perfect [38]. To stick to our "error-free" wish, we decide to craft for each source a dedicated article extractor based on the HTML structure analysis and BeautifulSoup. It removes CSS, scripts, and images. It removes commercial call to action.
- **Article Parser**: Extracting the title, the date of publication, the authors, the article text. It collects appendix at the bottom such as IOC section. It collects all reference links.

The columns of the Knowledge Database (KBDB) are source title, source url, article url, article title, article date, article authors, article text, article ioc section, article reference list and the article retrieval date.

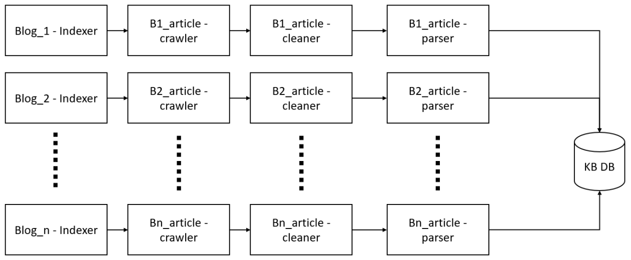

Figure 1 – Data Collector Inline Pre-processing

3.3 Entity Matcher

The Entity Matcher has two types of inputs: live knowledge from the KBDB and actualized reference framework. It produces as output an entity database.

3.3.1 Malware and Threat Actor name references

Malware references information is collected from the MITRE ATT&CK JSON (v6.3) as well as eight MISP Galaxies Cluster: Android, Banker, Botnet, Exploit Kit, Malpedia, Ransomware, RAT, Stealer.

Threat Actor references information is collected from the MITRE ATT&CK JSON as well as the Threat Actor MISP Galaxies Cluster.

3.3.2 Malware and Threat Actor name resolver

As there is no shared naming convention in the CTI industry regarding malware and threat actor, unlike in medical research, a same malware or threat actor may have several names across article publishers, and across time for the same publisher due to the abductive nature of investigation.

References sources such as MITRE ATT&CK or MISP Galaxies will use a name as a key index, and list related names respectively as aliases and synonyms (referred here as alias). In addition to the name variation, there is also the possibility of segmentation variation. A bag of MISP name and alias may be considered as two bags in ATT&CK, and vice-versa.

We merge by prioritizing the MITRE ATT&CK reference order.

3.3.3 Malware and Threat Actor name disambiguation

Some malware names or aliases and threat actor names or aliases are very generic, such as elfin, snake, silence, butterfly, biscuit, spaceship, or beta.

To stick to our "error-free" wish, we use ten books from the Gutenberg project [39] and three other long texts.

Arsene Kupin
Down and Out in the Magic Kingdom
In Story-land
Moby Dick
QA Panel at WWW2004
Stories of Useful Inventions
THE DISCOTECA PUBLICA MUNICIPAL DE SAO PAULO COLLECTION
The Hacker Crackdown
THE HAUNTERS & THE HAUNTED
The Hitchhikers Guide to the Internet
The World Bank - Justice Sector Institutional Strengthening Project
Trips to the moon
Wonder Tales from Many Lands

Table 2- Long texts list for whitelisting

All those long texts predate by long the modern cybercrime era or are from completely different topics. Each of their words are tokenized. If one of the malware or threat actor name or alias is in one of those texts, it means that it is probably ambiguous and may generate false positive.

We first check if the main name is flagged as ambiguous. If yes, we take its first alias that is not, and switch it as the main name. After this operation, each combo is flagged as 0 if none are ambiguous, 1 if the alias is ambiguous, and 2 if the name and the alias are ambiguous.

Those flags can be triggered during the later phases to choose the level of error which we are ready to work with.

3.3.4 MITRE ATT&CK techniques synonyms

Regarding the literature, it seems that the sheer volume of labelled data, the texts describing the technique's procedures, does not allow to create appropriate machine-learning text sequence recognition. The ~80% RoBERTa Large model accuracy from Alves is too low for us to handle regarding our article volume and too consuming for our computing resources.

To overcome the limitation highlighted by Legoy and Alves, we decided to use the same principle as for malware and threat actor: the main name is the ATT&CK technique name, and we create aliases based on selected keywords or sequences that evoke that technique. For example, we associate to "Application Shimming" the following aliases: shim cache, sdbinst.exe, or sysmain.sdb.

This also allows to capture a keyword or a sequence that encompasses several ATT&CK techniques. For example, the keyword "spearphishing" suggests "phishing", but also "Gather Victim Identity Information". It can also help to capture complex behaviour with basic keywords. For example, "each victim" suggests strongly "Execution Guardrails".

Using this process, we augment our techniques' pattern matching with 5855 keywords, each attached to the relevant MITRE ATT&CK techniques.

This solution may seem fastidious. It is. But for the same amount of time, annotating and labelling data to train machine-learning recognition would be less performing that a few keywords by techniques carefully selected by a CTI practitioner.

This solution may seem error prone. It is again. But we have here to remember that we are at the stage where we induce automated or human behaviour from forensic traces or malware analysis, that is abductive and inductive, and so error prone. Even the most perfect technical and automated solution would not be error proof as we are here at the interpretation level, and misjudgement, bias, or deviation are inherently part of the rules.

3.3.5 Geolocation data

To capture countries name and be able later to plot related information on a map, we collect on GitHub the Mohammed Le Doze's countries JSON file. From which we collect the latitude, the longitude, the main city name, the country's common name, the country official name, the country's region (e.g.: Africa) and the country's subregion (e.g.: Eastern Africa).

We manually added four region denominations: Middle East, Southeast Asia, Latin America, and Commonwealth of Independent States.

3.3.6 Entity Matcher

To perform our pattern matching operation, we rely on two tools: python's regex module and Spacy's PhraseMatcher with en_core_web_sm as model. The perimeter on which we apply them are one article's title and its related article text.

We use regex to match vulnerability expressed in the CVE format. We use Spacy's PhraseMatcher to catch the rest.

Each match for a given category is increasing a category counter. At the end of the process, each article has the list of sequences that match, the list of main names of the matched entities and a completeness score. Each category of entity (malware, threat actor, technique, CVE, geo area) gives one point if its counter is a least of one. Meaning that an article with a score of 5 mentioned at least one element in the five categories. An article with 0 produced no matching.

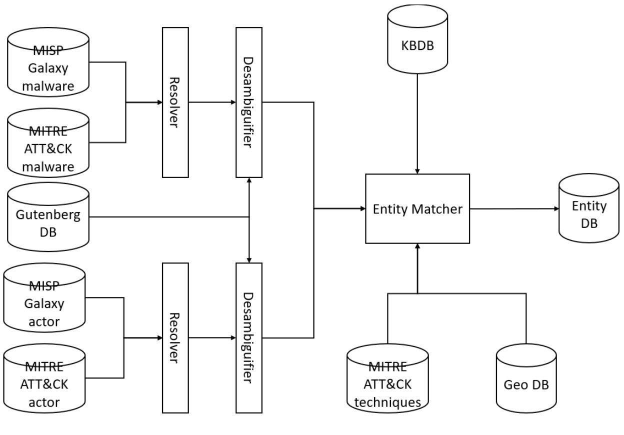

Figure 2 – Entity Matcher Components

The result of this processing is stored in the Entity Database. It contains the category of the entity, the matching phrase, the matching sequence, the matched entity, the reference of the source article and the completeness score.

3.4 Analytics Portfolio

The Analytics Portfolio component is a set of operational modules providing mission critical information.

3.4.1 QuantitativeAnalyzer

QuantitativeAnalyzer computes metrics about articles, metadata, and entities. It then projects them across time, geographical space or between themselves.

The goal is to produce dashboard and data sheet to perform further data mining tasks or data visualization. Some of the results are presented in the Results chapter.

3.4.2 EntityTrainer

Named Entity Recognition (NER), the ability to detect a known or unknown named entity in a text, mostly relies on supervised methods using large amount of high-quality annotated data [40].

One of the limits of our pattern matching approach is that it is dependable of the update of underlying references bases. CVE has a predictable form but are sometimes referred with names (e.g.: printnightmare, proxyshell, eternalblue). Geographic area names do not change regularly. We put aside technique recognition as Alves hints that it is more leaning toward sequence multiclass labelling tasks rather than named entity recognition.

Focusing on malware names and aliases, threat actor names and aliases, and vulnerability names, we leverage the Entity Database. We reuse the sentence, the entity category as entity label, the matched sequence as training entity. This allows to quickly create annotation files in IOB format and in the custom Spacy JSON format.

We also add another disambiguation step between malware and threat actor. In practice, an antivirus rooted article producer may use a malware name to name the associated threat actor (e.g.: the Sofacy malware leads Kaspersky to name the associated group Sofacy). When we find a match between a malware name or alias with a threat actor name or alias, we add several suffixes to the threat actor name or alias, such as "Group", "APT, or "Threat Actor".

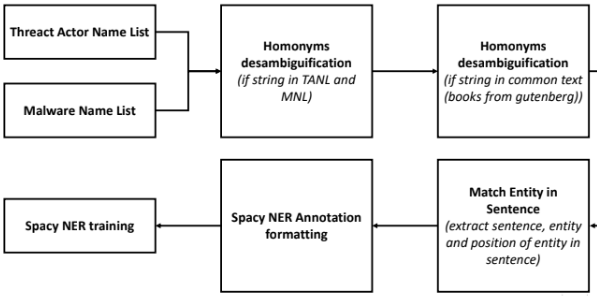
Threact Actor Name List
Malware Name List
Homonyms desambiguification (if string in TANL and MNL)
Homonyms desambiguification (if string in common text (books from gutenberg))
Match Entity in Sentence (extract sentence, entity and position of entity in sentence)
Spacy NER Annotation formatting
Spacy NER training

Figure 3 – Entity Trainer Analytics Workflow

This annotation dataset can be used to custom on-the-shelf NER modules, such as Spacy NER. This allows to catch malware and threat actor mentions that are not known from the references database. This allows to catch vulnerability mentions that are not in the CVE format. It introduces a greater risk of errors.

3.4.3 AbductionReductor

Based on our domain knowledge and the criminology literature review, we identified two notions that inspired us the AbductionReductor module. The first notion is the "habitual criminal". The criminal will use the same modus operandi over and over and change it only by necessity.

Postulate 4: Low-sophistication threat actors' ("Script Kiddies") modus operandi is bounded to its tool techniques implementation.
Postulate 5: Cybercriminal's modus operandi is bounded to their Return Over Investment (ROI). They will evolve their modus operandi only if it improves or saves the ROI.
Postulate 6: High-sophistication threat actors' ("APT") modus operandi will evolve following operational requirements but keep a structural base.

Based on these postulates and our experience, we can draw further postulates on the operational preparation and execution.

Postulate 7: A Threat Actor is composed of humans ("agent") that built personal habits that are difficult to break. Those habits are steered in time by the Threat Actor procedures ("agent's comfort zone technique set »").
Postulate 8: When breaking into a network ("a mission"), the Threat Actor has been given an objective.
Postulate 9: A mission requires the use of a minimum set of offensive techniques to fulfil the objectives that will be picked among the agent's comfort zone technique set ("primary technique set").

Postulate 10: The sum of agents' comfort zone never fully covers the mission requirements. The gap is the "secondary technique set".

Postulates 4 to 10 leads use to formulate our "Cyber Operation Constraint Principle".

Principle: "To fulfil a cyber offensive mission, a Threat Actor must execute a subset of offensive actions constrained by the disjointed subset of actions which it already masters."

The immediate implication of this principle is that there is a directed relationship and hierarchy in the offensive techniques used by a Threat Actor.

The second notion is the abductive nature of investigation. On which we can add the discretion or secrecy factor. When publishing a cyber intrusion report, without holding back any information, the author may have missed some part of the operation. Another extreme case is when publishing a cyber intrusion report, without missing any threat actor action, the author may hold back portion of its findings to preserve an edge over the competition or the adversary.

From a normalized database, based on the postulates and principle described, it is possible to apply an association algorithm to uncover the relationship strength between the use by choice of a master technique and the involved constrained use of a secondary technique. We chose the "A priori" algorithm because it was for us the easiest to implement in term of skills. We coded in-house because it was a fun exercise.

We start by selecting all articles with a single Threat Actor name and create a transaction table. One transaction is one article and a bag of MITRE ATT&CK techniques.

We then create the frequent itemsets table, from singleton to n-item. We have here created two modes: "full" and "grouped":

- The full mode computes the itemsets over the whole transaction table. The minimal support is of 0,2%, the minimal occurrence of a technique across all transaction.
- The grouped mode starts by creating sub-transaction table based on the Threat Actor name. The itemsets are computed over the sub-transaction tables. The minimal support is of 3, the minimal occurrence of a technique across all transaction.

Finally, we compute the Association Rules over the singleton. We compute the lift metrics and remove all rules less than 1. The lift metric allows to disregard "association by chance" just because the two elements are very common. It is useful for identifying which rules are most interesting and worth further exploration.

3.4.4 STIXCrafter

As we have extracted entities for each processed article, we could have translated them into the relevant STIX Domain Object and pack them together without relationship inside a STIX "Bundle SDO". But we wanted to go further and create meaningful relationship between SDO.

Our first operation is to approximate one article to a STIX Report but also to a STIX Campaign. CVE are translated as a Vulnerability SDO. We translate malware as a Malware SDO. We translate Threat Actor as Threat Actor SDO. We translate the techniques as Attack Pattern SDO. We translate geographical area as Location SDO.

Additional basic operations include:

- Creating a md5 hash of the title, adding it as a suffix of the Threat Actor Name and translating the string as an Intrusion Set.
- Creating an Identity SDO for each of our source.
- Creating a STIX Marking Definition Object to convey the authorship of the information.
- Creating a STIX TLP Marking Object to convey the diffusion level (here all as TLP-WHITE).
- Creating an Opinion SDO by using HuggingFace's transformers and the distilbert-base-cased-distilled-squad model, asking the question "What is the main opinion expressed in this text?". If the context (the article text) is two long, we use a recursive approach until its works (five rotation maximum): divide-and-conquer by cutting in half the sentences of the text, apply a summarization transformation using HuggingFace's transformers and the facebook/bart-large-cnn model, merge the summary and try again to answer the question.

To create our relationships, we take the semantic approach and the Subject-Verb-Object paradigm as explored by Usari. We will select the sentences where we have entity recognized by matching pattern and apply a standard Spacy NER operation using en_core_web_sm.

We focus on creating the following relationship:

- Victim Individual or Organization (Identity SDO): assuming an active sentence, semantically a victim is the object of a confrontational verb with a Threat Actor entity as subject. We select the object if it has been recognized by Spacy NER as a Person, a Group (NORP) or an organization. The associated SRO is "targets". We perform the relevant adaptation for passive sentences.
- Victim Region or Country (Location SDO): assuming an active sentence, semantically a victim is the Location entity object of a confrontational verb with a Threat Actor entity as subject. The associated SRO is "targets". We perform the relevant adaptation for passive sentences.
- Victim Technological Domain (Infrastructure SDO): assuming an active sentence, semantically a victim is the object of a confrontational verb with a Threat Actor entity or a malware entity as subject. We select the object if it has been recognized by Spacy NER as a Product or a Work of Art (yes, the pleasure of machine learning based matching). The associated SRO is "targets". We perform the relevant adaptation for passive sentences.
- Malware implemented techniques (Malware SDO): assuming an active sentence, the subject is a malware entity, the verb is part of in the usability lexicon and the object a technique. The associated custom SRO is "implements" following MITRE ATT&CK modeling, that diverges from the STIX original relationship which we do not agree with. We perform the relevant adaptation for passive sentences.
- Using the same principle, we developed further relationship classification.
- At the end, all attack patterns not linked to a malware are linked to the Intrusion Set. All malwares are linked to the Intrusion Set. All objects ID are references in the Report SDO (object_refs).

Source	Relationship	Destination
Threat Actor	Targets	Victim
Threat Actor	Targets	Location
Threat Actor	Targets	Technology
Malware	Targets	Technology
Malware	Implements	Techniques
Techniques	Targets	Vulnerability
Intrusion Set	Targets	Vulnerability
Malware	Exploits	Vulnerability
Technology	Has	Vulnerability

Table 3 – Implemented relationship using SVO paradigm and entity constraints.

Depending on the level of formalism we want in the final STIX JSON, we can apply the constraint over zero to three of the SVO combo.

4 Results

As the perimeter of the experiment is large, we have selected a few results to share.

4.1 Data Collector

We collected and processed 17153 articles, gathered from twenty-six sources, covering the 2007-10-27 / 2020-09-22 period.

Our malware reference base is composed of 2790 malware names and aliases. 2682 are classified as unambiguous (96.1%). Our Threat Actor reference base is composed of 851 adversary names and aliases. 823 are classified as unambiguous (96.7%). Our techniques reference base is composed of 535 adversary names and aliases. 263 have no additional keywords. Our location reference base is composed of 716 places.

4.2 Entity Matcher

4.2.1 Malware

Over our dataset, we match 14 886 malware names or aliases mentions (unambiguous) on 5 986 articles (34%).

Malware	Hits	%
Wannacry	321	2
Mirai	254	2
Dridex	245	2
Zbot	236	2
Emotet	233	2

Table 4 – Top 5 most cited malware

We counted 1201 unique mentions, meaning that 56% of the unambiguous reference base triggers no matching.

4.2.2 Threat Actor

Over our dataset, we match 4 915 threat actor names or aliases mentions (unambiguous) on 2 175 articles (12%).

Threat Actor	Hits	%
APT28	428	9
Shadow Brokers	260	5
Equation Group	229	5
Turla	227	5
Lazarus	181	4

Table 5 – Top 5 most cited Threat Actor

We counted 222 unique mentions, meaning that 57% of the unambiguous reference base triggers no matching.

4.2.3 Techniques

Over our dataset, we match 73 834 techniques mentions on 12 722 articles (74%).

Techniques	Hits	%
Obfuscated Files or Information	11 209	15
Command and Scripting Interpreter	7 398	10
Standard Application Layer Protocol	3 889	5
Spearphishing Attachment	2 870	4
Encrypted Channel	2 099	3

Table 6 – Top 5 most cited techniques

We counted 269 unique mentions, meaning that 50% of the reference base triggers no matching.

4.2.4 Location

Over our dataset, we match 24 334 country mentions on 3 369 articles (37%).

Techniques	Hits	%
United States of America	1 229	5
People's Republic of China	1 206	5
Federation of Russia	1 169	5
Europe	922	4
Germany	841	3

Table 7 – Top 5 most cited location

We counted 362 unique mentions, meaning that 50% of the reference base triggers no matching.

4.2.5 Vulnerability

Over our dataset, we match 12 085 CVE mentions on 2 856 articles (16%).

Vulnerability	Hits	%
CVE-2012-0158	114	1
CVE-2017-11882	88	1
CVE-2017-0199	66	1
CVE-2019-0708	66	1
CVE-2010-3333	44	0.3

Table 8 – Top 5 most cited CVE

4.2.6 Completeness score

Over the full dataset, the completeness score median is four.

Completeness score	Hits	%
4	10 619	61
3	2 699	16
5	2332	14
2	1154	7
1	349	2

Table 9 – Completeness score distribution

4.3 Analytics Portfolio

4.3.1 QuantitativeAnalyzer

We have 183 Threat Actors with bags of unique techniques they used at least once.

Technique	Hits	%
Obfuscated Files or Information	142	77
Command and Scripting Interpreter	127	69
Application Layer Protocol	118	64
Ingress Tool Transfer	102	55
Spearphishing Attachment	98	53

Table 10 – Top 5 most common techniques shared by Threat Actor

We have a median size of techniques bags of 21, with the maximal at 107 and the non-null minimal at 1.

Threat Actor	Size
Turla	107
Sandworm Team	105
APT28	103
Cobalt Group	100
Equation Group	99

Table 11 – Top 5 Threat Actors by biggest techniques bag

With this reference set we decided to perform some tests over recent publications with unattributed actors. We manually annotated the tested article into MITRE ATT&CK techniques (set B) and computed three metrics when compared with a given known threat actor set (set A):

- Jaccard coefficient (intersection divided by the union): metric of how the two sets are similar.
- Overlap coefficient (intersection divided by the size of the smallest set): metric of how the two sets are similar with an indication of how much the smaller set is inside the bigger set.
- Tversky Index (intersection divided by element of A not in B and element in B not in A): generalization of the Jaccard coefficient that is more robust to sets' size difference and outliers.

We finally aggregate those metrics with a formula inspired by the F1-score:

$$\text{Global Score} = \frac{2*J*O*T}{J+O+T}$$

A test has been performed on a fresh article from Securelist, titled “Bad magic: new APT found in the area of Russo-Ukrainian conflict”, dated 2023-03-21. This is not meant as an attribution but may orient investigators in a removal of doubt approach.

TTP	Jaccard	Overlap	Tversky	GS
TA505	0.21	0.65	2.14	0.19
Cutting Kitten	0.20	0.65	2.14	0.18
Invisimole	0.22	0.52	1.2	0.14
Molerats	0.20	0.52	1.2	0.13
APT37	0.19	0.52	1.2	0.12

Table 12 – Top 5 Threat Actor TTP overlap (by Global score)

4.3.2 EntityTrainer

We have been able to generate 84 656 training phrases for Spacy. On which we added 12 344 phrases without any mention and 3 000 empty phrases, leading to a 100 000-annotation set.

Entity Type	#
Malware	58 850
Threat Actor	17 581
Vulnerability	13 135
No entity mention	12 344
Empty phrase	3000

Table 13 – Breakdown of training phrases per entity label

4.3.3 AbductionReductor over the full dataset

The test of the AbductionReductor has been perfomed on a Transaction Table of size 12 808. The minimal frequent item support is 25, meaning that a technique must be at least in 25 transactions (articles) to be considered as a frequent singleton. The full process lasted for 5 minutes and 51 seconds. 3 088 frequent itemsets have been found, with n_{max} = 12.

Frequent n-itemset	#
1 (singleton)	133
2	1139
3	1928
4	3626
5	2997
6	2173
7	1423
8	770
9	312
10	86
11	14
12	1

Table 14 – Number of frequent n-itemsets

The longest 12-itemsets occurred in 38 articles (0.39%): Obfuscated Files or Information, Command-Line Interface, Scripting, Remote Desktop Protocol, PowerShell, Replication Through Removable Media, Process Hollowing, Remote File Copy, Screen Capture, Deobfuscate/Decode Files or Information, Exploitation of Remote Services, and Virtualization/Sandbox Evasion.

The results of the Association Rules are 1434 associations, score with the antecedent technique support (how many times it appears in the transaction table), the consequent technique confidence (how many times it appears in the transaction where the antecedent technique appears) and the lift metrics (how many times the tuple appears divided by the sum of the support of each technique).

Antecedent Technique Consequent Technique	Sup.	%Conf	Lift
Command-Line Interface System Network Connections Discovery	12	68	349
Virtualization/Sandbox Evasion Process Hollowing	11	65	349
Command-Line Interface System Hidden Window	12	66	334
Standard Application Layer Protocol Dynamic DNS	29	98	321
Exploitation of Remote Services Data from Information Repositories	14	69	318

Table 15 – Top 5 Association Rules (by Lift)

We can also look at the following techniques by their highest support and highest lift for each of their first occurrence.

Antecedent Technique Consequent Technique	Sup.	%Conf	Lift
Obfuscated Files or Information Data Encoding	43	96	206
Scripting Local Job Scheduling	41	89	188
Standard Application Layer Protocol Dynamic DNS	29	96	247
Spearphishing Attachment Template Injection	17	60	209
Remote File Copy Local Job Scheduling	15	68	285

Table 16 – Top highest support and highest lift Association Rules for unique antecedent techniques

4.3.4 AdbuctionReductor on Threat Actor subset

We performed the same operation on subset of articles based on the threat actor. 16 do have at least one association rule over the 222 Threat Actor unique mentions (7%).

Threat Actor	#Rules	Median Sup.	Median %Conf	Median Lift
APT28	205	20	71	1.31
Sandworm Team	138	14	60	1.47
Equation Group	110	17	62	1.40
Turla	103	14	66	1.31
Lazarus	100	10	66	1.57

Table 17 – Top 5 Threat Actor by number of Association Rules

The longest APT28's itemset occurred in 4 articles (5.71%) and is of size nine: Obfuscated Files or Information, Scripting, Security Software Discovery, Remote File Copy, Screen Capture, Deobfuscate/Decode Files or Information, Spearphishing Attachment, System Information Discovery, Standard Application Layer Protocol.

The most frequent APT28's non-singleton itemset occurred in 28 articles (40%) and is of size two: Obfuscated Files or Information, and Scripting.

APT28's rules support has a median of 20, the median confidence is 66 and the median lift is 1.36.

Antecedent Technique **Consequent Technique**	**Sup.**	**%Conf**	**Lift**
Custom Cryptographic Protocol Logon Scripts	12	100	5.44
Modify Registry Registry Run Keys / Start Folder	8	80	5.23
Rundll32 Logon Scripts	8	75	4.59
Software Packing Windows Management Instrumentation	8	75	4.59
Process Discovery Peripheral Device Discovery	20	100	3.5

Table 18 – Top 5 APT28's Association Rules (by lift)

In comparison, Equation rules support has a median of 17, the median confidence is 62 and the median lift is 1.40.

Antecedent Technique **Consequent Technique**	**Sup.**	**%Conf**	**Lift**
Input Capture Clipboard Data	17	100	3.74
Remote Access Tools File and Directory Discovery	11	80	3.59
Credential Dumping Clipboard Data	13	80	3.19
Software Packing Execution through Module Load	11	75	3.16
Execution through API Execution through Module Load	22	100	2.99

Table 19 – Top 5 Equation's Association Rules (by lift)

4.3.5 STIXCrafter

For the victim linker, we conducted quality benchmark on our relationship extraction and classification rules. The test consisted of 86 true positive documents and 50 true negative documents. At the time of the test, we only used the SVO approach without entity constraint. The first metric displays the detection of a victim in a document or not. Not the qualitative identification of the victim. The second does.

	Has Victim	**No Victim**
Victim Detected	63 (TP)	20 (FP)
No detection	23 (FN)	30 (TN)

Table 20 – Victim Detection Confusion Matrix

Further metrics are the following: Precision: 75%, Recall: 75%, and F1-Score: 75%.

	Victim	**No Victim**
Correct Identification	59 (TP)	20 (FP)
No or incorrect identification	27 (FN)	30 (TN)

Table 21 – Victim Correct Identification Confusion Matrix

Further metrics are the following: Precision: 68%, Recall: 66%, and F1-Score: 67%.

5 Discussion

As the perimeter of the experiment is large, we have selected a few topics to discuss on.

5.1 Background and related work

We acknowledge that our literature review is not complete and does not encompass the full spectrum of the topics that we cover.

Yet, we set some cornerstones to the practice of Cyber Threat Intelligence by stating its abductive and inductive nature. Our contribution is to take, in addition to the nature of investigation, editorial censorship unlike previous research [41][42]. We also emphasize the strong link that it has with criminology and show that TTP or modus operandi narrative normalization for crime prevention and detection is a century old problem. Future research could approach the problem from both disciplines, as solving one will very certainly impact positively the other.

Finally, thanks to classical criminology theory, we have been able to formulate our Cyber Operation

Constraint Principle. We believe that it can change the perspective of defender on how the threat actor is evolving its modus operandi, if it can change at all.

We are convinced that in addition to the two notions we based our principle on - the routine activity theory and the abductive nature of investigation - a third one exists: not only the threat actor is constrained by its own habits and its group's habits, but the environment does also limit its ability to change freely its TTP. Because computerized systems have a finite set of possibilities to perform a task, such as downloading a file from the internet. System call, API, and all layers from the electronic to the application data layer, do introduce logical and structural constraints. We have yet been unable to find the right literature to support this assertion.

Finally, the formulation of the Cyber Operation Constraint Principle leads us to create a tool, the AbductionReductor. Not only was it very fun to develop, but the output is surprisingly good given our expectation and stimulating from a domain centric analytical perspective. Notably for the result of the Threat Actor subset streams, that we expected to be too small to give any meaningful result. The apriori algorithm is probably not the most computing optimal, but our volume of data is reasonable. Even with our outdated dataset, we are still today playing with the AbductionReductor to gather insight and explore intriguing correlation.

5.2 Data Collector

The way we choose to collect carefully and specifically each source forced us to continuously maintain and monitor the HTML pages structural changes in source. This is an overhead on the automation but that we are ready to accept as we have not found easy alternatives for near-perfect article and metadata extraction.

Obvious limitation is the scalability of the solution, but this drawback is reduced to the limited amount of relevant source and the manageable volume of data to ingest every day, in comparison with other domains or tasks. As we also care about the quality of analysis we feed our system with, we believe it is a reasonable limitation.

Another limitation is when a source enforces anti-crawling techniques. Even if a bypass solution does exist and work well, this poses an ethical problem. Indeed, if it seems reasonable to bypass such protection for research purposes, it is another story if the derived data are meant to be included in a commercial offer. The community could share its ethical or lawful insight to advance the debate.

5.3 Entity Matcher

We have been agreeably surprised by our pattern matching approach and were not expecting such honorable results.

Regarding the location and the actor's metrics, we can suspect a tropism of publications towards USA related geopolitical agenda. This can be explained by our source selection that is not enough distributed, but not only. The USA is still a dynamic and profitable market and provider in terms of cybersecurity and talking in echo with them (it does not necessarily mean to be aligned with) is not a stupid idea business wise. Also, we must acknowledge that the world agenda, even more recently, is paced around the USA, The Federation of Russia, and The People's Republic of China.

Comparing malware metric and Threat Actor metric, we can see a pattern. Top 5 malwares are cybercrime related and Top 5 Threat Actors are APT related. This may indicate that cybercrime discussions are focusing on the tool (the "how"), while sophisticated cyber-attack discussions are focusing on the perpetrator (the "why" and the "who"). It is quite interesting, because there is probably more chances to catch a cybercriminal individual than a cyber-spy or cyber-mercenary. It may also indicate that it is easier to spot cybercrime related malwares, as spotted targeted malwares necessarily require a vast and diversified collection pool, such as the commercial distribution of endpoint protection. One big limitation here is that since the end of the article collection, ransomware gangs heavily hit the news. It would be interesting to see if some of them made it to the Threat Actor Top Five during the last years.

Looking at top techniques' statistics, we can immediately see the gap in detection number, even inside the top 5. We can draw here two hypotheses: the most detected techniques by our system are the ones with simple and discriminatory keywords (e.g.: obfuscation). The second hypothesis is that not only are they simpler for our system to spot, but they are also very visible and common techniques in the intrusion process. Investigators cannot miss them. More vicious or complex techniques would probably take a full paper to explain and could not be grasped in a narrowed text sequence nor keyword.

We would have expected more vulnerability mentions. Our identification is probably hindered by our sole reliance on CVE schema matching. Also, we can note two vulnerabilities in the top 5 that are more than five-year-old at the time of the end of the collection. As we have not presented this top with a chronological context, it is hard to draw any conclusion or observation. Future research should be done to look at their distribution of occurrence over time.

Finally, we were very surprised that more than 60% of collected articles include at least four of the five entities we were focusing on. We were expecting a lower completeness score.

5.4 Analytics Portfolio

5.4.1 QuantitativeAnalyzer

Our approach of comparing bags of techniques has been developed following previous unpublished work based on IoC, used until 2017. We would take a bag of indicators and metadata of a fresh investigation or an unattributed operation and compare it to a knowledge base set of indicators attributed to a Threat Actor. We applied this approach to unattributed modus operandi and attributed modus operandi. We were expecting mixed and modest results as the size of the sets are very limited.

Running our test on various articles shows interesting results. The one provided in the Results chapter shows interesting results (such as the recent Invisimole activity in Ukraine) considering the abductive nature of investigation and our limited visibility on the details of the - to date - unattributed Bad magic event. Obvious limitation is our outdated dataset. Also, this approach just looks at individual technique matching. We are currently exploring how to combine this similarity analysis with the AbductionReductor tool.

5.5 Entity Trainer

We have not used nor tested our Entity Trainer annotation dataset. Future research may include it in our entity recognition process.

5.6 AbductionReductor

The premises of the Cyber Operation Constraint Principle and the vision of AbductionReductor were the starting point of engaging this whole NLP work. The very first version was only based on links between Threat Actors and techniques stored in MITRE ATT&CK. The Threat Actor was the transaction, instead of the article as of today. The results were disappointing of course.

Qualitatively reviewing the long itemsets and the association rules, computed over the full dataset, with a penetration tester and a Red Teamer, was a fun and interesting exercise. Even if some combination still puzzled us, the discussion among strong association rules oscillated between obvious acknowledgement and positive curiosity.

Our greatest positive surprised was with the threat actor segmentation approach, that we were expecting to be of low interest because of the little amount of value. But we were wrong. Even if only 16 threat actors have association rules, based on our settings, the association rules review for top threat actors were surprisingly insightful with offensive security specialists. We were able to, if not prove, acknowledge the possible divergence in modus operandi based on our “primary technique”/ ”secondary technique” (Postulates 9, 10).

Yet, a limitation in interpretation of those co-occurrences is that we cannot at the moment know which of the antecedent technique / consequent technique is the primary technique (inside the threat actor comfort zone) or the secondary technique (outside the threat actor comfort zone and inside the mission requirements).

The implication for the forensic investigation side is that using a tool such as AbductionReductor allows to change the research methodology from an exploratory and instinct based approach to a confirmatory or assisted approach. Cyber Threat Intelligence Analyst or Threat Hunter could use it for inspiration or correlation. In the context of withheld information for competitive advantage, the tool could fill the holes of the editorial censorship.

On the other hand, the offensive side could use the tool to design missions. Either to “desilhouetting” (the ability to conceal its appearance) and degrade modus operandi-based attribution hints. Or to create missions with similar behaviour to other modus operandi, but still rooted in their primary technique set.

Our implementation is still crippled with pending optimization and limitation, but we notice that the techniques class imbalance, due to our biased recognition system and the biased underlying security researchers’ publications, does not completely hinder the ability to generate insightful combination.

5.7 STIXCrafter

We had very little expectation on this first approach and were positively surprised. The linguistic based approach is interesting in our "error-free" wish as we can carefully fine-tune what we want to detect. The immediate limitation is the time-consuming process.

The main limitation is our performance metrics. As we are not data mining or data science specialists, as we did not create this project with benchmarking in mind, they should be considered with caution. In any case, it seems way less performing than related approach [43][44].

The qualitative review of false positive, on the false positive document set, was the most interesting part. Most of our matches were indeed victims of an attack or infection. But not in our domain. Some examples of false positive:

- Locust invasion in Latin America corn fields.
- A Guatemalan bacteria infecting Michigan geranium.
- The effects on coughing in an airplane.
- An attack of a WHO vehicle in Myanmar.
- An elderly murder victim.

We see those false positives as the product of a shared semantic between criminology and infection disease and this reinforces our conviction that advanced or best practice in one field in the detection of prevention may lead to improve the other. If NLP tasks are facilitated in medical research, it is probably because of a large public and private research sector, that produces very structured documentation and shares a common language. While it is understandable for the cybersecurity private sector not to engage in this way, it is a call to produce more empirical and structured academic research on cyber-attacks, malware behaviour and threat actor modus operandi.

5.8 General reflexion on NLP

During the active development of the project, between 2019 and 2021, there was not a month without a breakthrough in the NLP field. The field is very stimulating, oscillating between linguistics and computer science.

The main challenges that we encounter in Natural Language Understanding were class imbalance, homonymy, relation classification, coreference and anaphora, diachrony and synchrony. We trust that specialists and researchers will find obvious existing solutions to improve our system or develop custom adaptations. We also trust that the advance in and the accessibility of Large Language Model (LLM) such as the GPT3+, Chinchilla, LLaMA or LaMDA can tackle most of our problems.

We believe that diachrony and synchrony of domain specific language is the key challenge which future research should focus on, deeply rooted in linguistics. The language evolves with time. Whether it be the structure of the phrase or the wording itself. In the 1990's we would have written about "computer viruses" when now we talk about malwares. Even the concept of a malware type evolves, as in the 1990's the structure of a software was more turned to monolithic and homogenous functionalities, whereas today it is more modular and with agile functionalities evolution through frameworks. When was spearphishing coined for the first time? Does it mean that before that, it did not exist? In our articles collection, ranging from 2007 to 2020, the language evolves and yet we took a synchronic approach, based on a reference base of our time. Even the reference base does evolve, such as MITRE ATT&CK, in version 6.3 when we included it in our system, and now in version 12.1, with new framing, naming, phrasing. In the hypothesis of reactivating the flow of articles, should we keep the 6.3 with the risk of missing precision in current modus operandi recognition, or should we update losing our precision on past events? The problem would stay the same with LLM, looking at the history of cyber-attacks only from a synchronic perspective. Future research on the application of NLP to cyber threat reports should look at the implication and the benefits of each approach. As modern conflicts are partly conducted through zero and one, we are talking about the ability for future generation to recollect and write The History.

Class Imbalance, regarding sources, techniques, actors, or any entity is deemed for us to be an inherent part of the application context. We urge future research not to search to reduce class imbalance. It must deal with it. Due to the abductive nature of cyber-attack investigation, the difference in activity intensity of actors, the effort they put (or do not put) to stay discreet and the world agenda. Class imbalance and missing data is a natural part of criminal activity investigation and analysis.

Homonymy between threat actor names, between malware names and between malware and threat actor names, is also deemed for us to last. We are in the context of clandestine activity investigation, performed by public institutions or private commercially motivated organizations. The sensitivity and the commercial context explain the siloed investigation effort. Also, each investigative organization with a collection capability does not have the same distribution of sensors, either geographically, in the typology of their constituencies or customers. This means that they may work on the same operation or threat actor, with non-overlapping clues, preventing them to tie this activity with a previous one. This is the difficult art of attribution and naming of clusters of activity.

Relation classification, coreference and anaphora – singly or all combined altogether - were frustrating. Yet, we believe that these challenges can be more easily overcome or reduced thanks to technological advance, including current LLM. Some tests using GPT3.5 showed promising solutions.

We have been way more circumspect when it comes to GTP3.5 performance at identifying and linking narrative modus operandi description segments to the relevant MITRE ATT&CK techniques. But it is very good at creating STIX2.1 JSON formatted information from an CTI report, if you don not care about the relationship.

5.9 The Cyber Operation Constraint Principle

The Cyber Operation Constraint Principle was a set of postulates in our head for a time. Thanks to our literature review, we were able to root it in the criminology field.

Even if it is for us the greatest achievement of this project, we need to dig deeper to reduce the set of postulates and find ties with the existing set of knowledge. Our first research leads lay in Computer Science, Cybernetics, Cognitive Science and Organizational theory. This would also allow to reframe or enrich the principle.

We can also use it as a foundation for future research involving intrusion path identification and analysis.

5.10 Key research limitations

The main limitation of our project lies in the lack of academic rigor when conducting it, which is impacting the exact reproducibility of our experiment. Also, the literature review may be incomplete.

As Cyber Threat Intelligence and Offensive Security specialists, we do not have the necessary background in data, software, and infrastructure engineering execution excellence that specialists of those domains would have. It is the same when it comes to data science and linguistics.

6 Conclusion

In this paper, we presented a system that addresses the lack of standardization and structure in Tactical Cyber Threat Intelligence (CTI). Our system collects blog articles about cyber-attacks, normalizes them with a defined vocabulary, and stores them in a structured language. We believe that this system has the potential to significantly improve the ability of security researchers to compare threat actor modus operandi.

Furthermore, we formulated a Cyber Operation Constraint Principle that could inform future research in Cyber Threat Intelligence. We derived from it a tool, the AbductionReductor, that we believe has the potential to augment partial knowledge about a threat actor's behaviour while investigating its actions. We also build a bridge between cyber adversary modus operandi analysis and classical criminals’ modus operandi analysis for crime prevention and detection.

Our system and principle could help reduce the workload of CTI analysts and provide more accurate and reliable intelligence for incident response and attack investigation. Our work also contributes to the emerging field of computational criminology, which aims to apply computational and data-driven methods to the study of crime and security.

We also identified key limitations of our project, including the lack of academic rigor and potential gaps in the literature review due to our specialized background as Cyber Threat Intelligence and Offensive Security specialists.

By further refining and enriching our principle, we hope to contribute to the growing body of knowledge on cyber operations and to the nascent field of computational Cyber Threat Intelligence. We also hope to contribute, by the communicating effect, to the global prevention and detection of crime.

We strongly advise the research community to develop methods that deal with class imbalance, not suppress it, as it is a natural parameter of investigation, to produce solutions usable by practitioners. We urge the research community to assess the challenges and benefits of synchrony and diachrony linguistics analysis. The stake is the ability for future generation to recollect and write “cyber history”.

Acknowledgment: This research received no funding or help.

Author details

Ronan Mouchoux

XRATOR
ronan@x-rator.com

François Moerman

XRATOR
francois@x-rator.com

References

[1] SCHLETTE, Daniel, CASELLI, Marco, et PERNUL, Günther. A comparative study on cyber threat intelligence: the security incident response perspective. IEEE Communications Surveys & Tutorials, 2021, vol. 23, no 4, p. 2525-2556.

[2] CHISMON, David et RUKS, Martyn. Threat intelligence: Collecting, analysing, evaluating. MWR InfoSecurity Ltd, 2015, vol. 3, no 2, p. 20-22.

[3] DOERR, Christian. Cyber Threat Intelligences Standards–A High Level Overview. TU Delft CTI Labs, 2018.

[4] BERADY, Aimad. Understanding sophisticated threats. 2022. Thèse de doctorat. CentraleSupélec.

[5] KWIATKOWSKI, Ivan, MOUCHOUX, Ronan, Automation and structured knowledge in Tactical Threat Intelligence. In : BotConf. 2018.

[6] RAHMAN, Md Rayhanur, MAHDAVI-HEZAVEH, Rezvan, et WILLIAMS, Laurie. A literature review on mining cyberthreat intelligence from unstructured texts. In : *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2020. p. 516-525.

[7] BRIDGES, Robert A., HUFFER, Kelly MT, JONES, Corinne L., *et al.* Cybersecurity automated information extraction techniques: Drawbacks of current methods, and enhanced extractors. In : *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017. p. 437-442.

[8] SOLLACI, Luciana B. et PEREIRA, Mauricio G. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of the medical library association, 2004, vol. 92, no 3, p. 364.

[9] WORLD HEALTH ORGANIZATION, et al. World Health Organization best practices for the naming of new human infectious diseases. In : World Health Organization best practices for the naming of new human infectious diseases. 2015.

[10] SKONIECZNY, Stanislaw. The IUPAC rules for naming organic molecules. Journal of chemical education, 2006, vol. 83, no 11, p. 1633.

[11] NWOGU, Kevin Ngozi. The medical research paper: Structure and functions. English for specific purposes, 1997, vol. 16, no 2, p. 119-138.

[12] HELBICH, Marco, HAGENAUER, Julian, LEITNER, Michael, et al. Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach. Cartography and Geographic Information Science, 2013, vol. 40, no 4, p. 326-336.

[13] GOODCHILD, Michael F. Citizens as sensors: the world of volunteered geography. GeoJournal, 2007, vol. 69, p. 211-221.

[14] CHEN, Hsinchun, CHUNG, Wingyan, XU, Jennifer Jie, et al. Crime data mining: a general framework and some examples. computer, 2004, vol. 37, no 4, p. 50-56.

[15] BIRKS, Daniel, COLEMAN, Alex, et JACKSON, David. Unsupervised identification of crime problems from police free-text data. Crime Science, 2020, vol. 9, no 1, p. 18.

[16] FOSDICK, Raymond B. Modus operandi system in the detection of criminals. J. Am. Inst. Crim. L. & Criminology, 1915, vol. 6, p. 560.

[17] CORNISH, Derek B. The procedural analysis of offending and its relevance for situational prevention. Crime prevention studies, 1994, vol. 3, no 1, p. 151-196.

[18] COHEN, Lawrence E. et FELSON, Marcus. Social change and crime rate trends: A routine activity approach. American sociological review, 1979, p. 588-608.

[19] RATCLIFFE, Jerry H. Intelligence-led policing. Routledge, 2016.

[20] CHAU, Michael, XU, Jennifer J., et CHEN, Hsinchun. Extracting meaningful entities from police narrative reports. 2002.

[21] SPENCER, W. Dean. Software development for AN. Georgia Institute of Technology, 1979.

[22] EDWARDS, Charles, MIGUES, Samuel, NEBEL, Roger, et al. System and method of data collection, processing, analysis, and annotation for monitoring cyber-threats and the notification thereof to subscribers. U.S. Patent Application No 09/950,820, 28 mars 2002.

[23] LOCARD, Edmond. Traité de criminalistique. J. Desvignes, 1931.

[24] CULLEY, Adrian. Computer forensics: past, present and future. Information security Technical report, 2003, vol. 8, no 2, p. 32-36.

[25] OATLEY, Giles, CHAPMAN, Brendan, et SPEERS, James. Forensic intelligence and the analytical process. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, vol. 10, no 3, p. e1354.

[26] EVANS, Jonathan St BT et OVER, David E. Reasoning to and from belief: Deduction and induction are still distinct. Thinking & Reasoning, 2013, vol. 19, no 3-4, p. 267-283.

[27] STROM, Blake E., APPLEBAUM, Andy, MILLER, Doug P., et al. Mitre att&ck: Design and philosophy. In : Technical report. The MITRE Corporation, 2018.

[28] ZHU, Ziyun et DUMITRAŞ, Tudor. Featuresmith: Automatically engineering features for malware detection by mining the security literature. In : Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016. p. 767-778.

[29] HUSARI, Ghaith, AL-SHAER, Ehab, AHMED, Mohiuddin, et al. Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources. In : Proceedings of the 33rd annual computer security applications conference. 2017. p. 103-115.

[30] HUSARI, Ghaith, NIU, Xi, CHU, Bill, et al. Using entropy and mutual information to extract threat actions from cyber threat intelligence. In : 2018 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2018. p. 1-6.

[31] AYOADE, Gbadebo, CHANDRA, Swarup, KHAN, Latifur, et al. Automated threat report classification over multi-source data. In : 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE, 2018. p. 236-245.

[32] THEIN, Thin Tharaphe, EZAWA, Yuki, NAKAGAWA, Shunta, et al. Paragraph-based estimation of cyber kill chain phase from threat intelligence reports. Journal of Information Processing, 2020, vol. 28, p. 1025-1029.

[33] LIN, Ling-Hsuan et HSIAO, Shun-Wen. Attack Tactic Identification by Transfer Learning of Language Model. arXiv preprint arXiv:2209.00263, 2022.

[34] LEGOY, Valentine Solange Marine. Retrieving att&ck tactics and techniques in cyber threat reports. 2019. Thèse de maîtrise. University of Twente.

[35] LEGOY, Valentine, CASELLI, Marco, SEIFERT, Christin, et al. Automated retrieval of att&ck tactics and techniques for cyber threat reports. arXiv preprint arXiv:2004.14322, 2020.

[36] YODER, Sarah. Automating mapping to att&ck: the threat report att&ck mapper (tram) tool. 2019.

[37] ALVES, Paulo MMR, GERALDO FILHO, P. R., et GONÇALVES, Vinícius P. Leveraging BERT's Power to Classify TTP from Unstructured Text. In : 2022 Workshop on Communication Networks and Power Systems (WCNPS). IEEE, 2022. p. 1-7.

[38] NWALA, Alexander. A survey of 5 boilerplate removal methods. 2017.

[39] STROUBE, Bryan. Literary freedom: Project gutenberg. XRDS: Crossroads, The ACM Magazine for Students, 2003, vol. 10, no 1, p. 3-3.

[40] LI, Maolong, YANG, Qiang, HE, Fuzhen, et al. An unsupervised learning approach for NER based on online encyclopedia. In : Web and Big Data: Third International Joint

Conference, APWeb-WAIM 2019, Chengdu, China, August 1–3, 2019, Proceedings, Part I 3. Springer International Publishing, 2019. p. 329-344.

[41] BROMANDER, Siri, JØSANG, Audun, et EIAN, Martin. Semantic Cyberthreat Modelling. STIDS, 2016, p. 74-78.

[42] HETTEMA, Hinne. Rationality constraints in cyber defense: incident handling, attribution and cyber threat intelligence. Computers & Security, 2021, vol. 109, p. 102396.

[43] JONES, Corinne L., BRIDGES, Robert A., HUFFER, Kelly MT, *et al.* Towards a relation extraction framework for cyber-security concepts. In : *Proceedings of the 10th Annual Cyber and Information Security Research Conference*. 2015. p. 1-4.

[44] RAHMAN, Md Rayhanur et WILLIAMS, Laurie. From Threat Reports to Continuous Threat Intelligence: A Comparison of Attack Technique Extraction Methods from Textual Artifacts. *arXiv preprint arXiv:2210.02601*, 2022.

PAGE LEFT BLANK