Toward transparent and trustworthy attribution of cyberthreats

1st Matěj HAVRÁNEK
Faculty of Information Technology
Czech Technical University
Prague, Czech Republic
matej.havranek@fit.cvut.cz

2nd Peter KÁLNAI
ESET Research
Prague, Czech Republic
0000-0002-1130-9180

3rd Hana KUBÁTOVÁ
Faculty of Information Technology
Czech Technical University
Prague, Czech Republic
hana.kubatova@fit.cvut.cz

***Abstract*—Cyber threat attribution is a critical component of modern cybersecurity, involving the identification of actors behind malicious activities and is defined as the process of identifying the origin, perpetrators, and intent behind a cyberattack. Its importance lies in enabling informed responses, deterring future attacks, and fostering accountability. It is a very dynamic field that faces significant challenges due to the unstable and rapidly evolving nature of cyber threats.**

In this research, we explore the balance between exactness and verifiability of cyber threat attribution. We define the basis for structuring data from cybersecurity incidents and a methodology for building a formal attribution based on this data, focusing on both of these aspects, which is intended to serve as the starting step to building a robust system for supporting formal attribution.

***Index Terms*—attribution, cyber threat, cyberattack, advanced persistent threat, malware, methodology**

I. State of the art

Cybernetic attribution is defined in [1] as the task of identifying the party that should be held politically responsible for an offensive cyber operation. The exactness of an attribution means precision in identifying and attributing specific actors or threats. Verifiability of an attribution denotes the ability to provide clear and reproducible evidence supporting the attribution claim, allowing a third party to reproduce the attribution based on the published evidence. In theory, there exist numerous formalizations of the attribution process, as well as frameworks and guidelines. In practice, however, attribution is still often done on an arbitrary basis and lacks formal methodology that can be used to verify its claims.

Current attribution methodics

Currently there exist multiple methodics for cyber threat attribution, ranging from more theoretical ideas to frameworks that have strong practical use. Some focus only on the use of malware by the attacker, while others provide a more complex overview of the entire malicious activity. Each method has a slightly different purpose and is suitable for a specific type of workflow and goals. The research in this paper is mainly based on the Diamond model [2], the Q model [3] and the MICTIC framework [4], each one combining multiple dimensions of the underlying data to form a comprehensive conclusion based on the connections between individual artifacts. Recently, there have been attempts to utilize artificial intelligence and machine learning approaches in attribution or helping attribution done by humans [5], but to our knowledge there is no established attribution methodic relying on these technologies - in all studied cases, these technologies are used to enhance the human-lead attribution process following one of the existing methodics or not relying on any methodics whatsoever. Despite the availability and relative ease of use of the above-mentioned methodics, a large part of attribution is still done based on individual beliefs and done without presenting relevant supporting evidence. This approach is called faith-based attribution [6].

II. Challenges in attribution

Attributing cyber threats is inherently difficult due to limited, often incomplete evidence. Investigations typically rely on indirect data — from automated systems, forensic reports, and malware analyses — each layer increasing the risk of misinterpretation or error. Data integrity issues and potential tampering further complicate this process.

A major obstacle is the lack of comprehensive, publicly available datasets to compare TTPs (tactics, techniques, and procedures) and IoCs (indicators of compromise), undermining the verifiability of attribution claims.

Practical application of attribution theory is also fragmented. Legal, technical, and procedural differences across organizations and nations hinder standardization. Attribution often depends on confidential or proprietary data that cannot be publicly disclosed — such as customer telemetry in private firms or classified intelligence in state agencies—making conclusions difficult to verify and heavily reliant on the authority of the attributor.

Balancing exactness and verifiability

Ideally, attribution would rely on accurate, reliable, and transparent data, ensuring both precision and verifiability. In practice, however, data limitations and confidentiality often force a trade-off between the two.

Prioritizing exactness typically means withholding evidence that cannot be publicly disclosed, making the attribution less verifiable and reliant on trust in the source. This is common

978-3-903176-74-4

in the industry, where confidence is placed in the attributor rather than shared data [7].

Conversely, emphasizing verifiability can reduce precision, as only shareable evidence is used. This cautious approach is typical in legal settings, where attribution must withstand external scrutiny, such as in criminal investigations.

III. Methodology

Let us try to formalize the attribution process. First, we start with some definitions.

- A *cyberattack is an event that occurs within the 7 steps of the cyber kill chain as introduced by Lockheed Martin [1].*
- *An* indicator of compromise (IoC) is a piece of digital evidence that witnesses the attacker's activity. We can view IoCs as outputs - raw sequences of bytes - from digital forensic software.
- A *collection of a cyberattack is a set of IoCs collected from a cyberattack, illustrated in Fig. 1.*
- *An* activity cluster (AC) is an abstract entity that performs a cyberattack. This may be a threat actor or one of their operations. Even a single cyberattack can be viewed as an AC. This is shown in Fig. 1.
- A *threat landscape is an undirected graph whose vertices are ACs, which are unconnected or connected with two types of edges labeled "different from" and "said to be the same as", as shown in Fig. 2.*

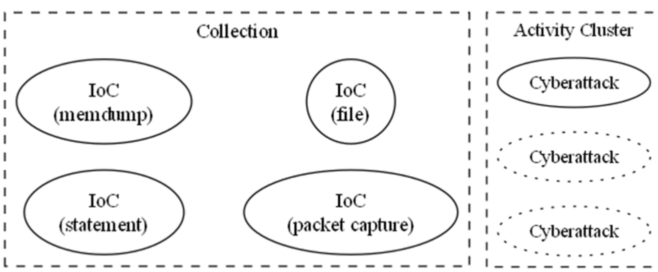

Fig. 1. A collection made of individual IoCs and an activity cluster made of one or more cyberattacks.

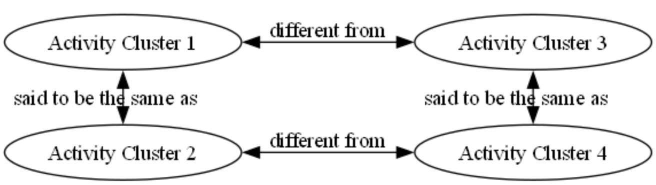

Fig. 2. An example of a threat landscape consisting of four ACs.

A collection may include a malicious file, memory dump, or network capture. However, more information can be leveraged in the forensic process, for example by analyzing data or applying various hashing algorithms, deobfuscation and extraction of samples, recognizing various high-level patterns, etc. For example, the statement *The attackers used the RC4 algorithm in one of the executables* is information that is not straightforward and often requires strenuous work to incorporate.

In order to model the data from the attribution process, we use a form of subject-predicate-object expressions, so-called RDF triples, [8], introduced in 1999 by W3C in order to store data in a machine readable way. We introduce three notions (item, property, and value) that correspond to the roles in the RDF triple. A property and a value form a statement.

An item is a basic element in the hierarchy of our structured data. Each item contains a set of statements related to it. A statement is a claim that holds true about an item, consisting of a property and its value. A property expresses a relation between - or a connection of - the item and its value. The value of a property can be one of the following categories: `Item`, `DateTime`, `String`, `Quantity`. The list of properties is finite, but eventually increases as we progressively model the attribution data. There are two basic membership properties with the "Item" type, *"instance of"* or *"subclass of"*. To show an example, *"SHA-1"* is an instance of *"cryptographic hash function"*, and *"cryptographic hash function"* is a subclass of *"algorithm"*. An example visualisation of a cyberattack with its items is shown in Fig. 3.

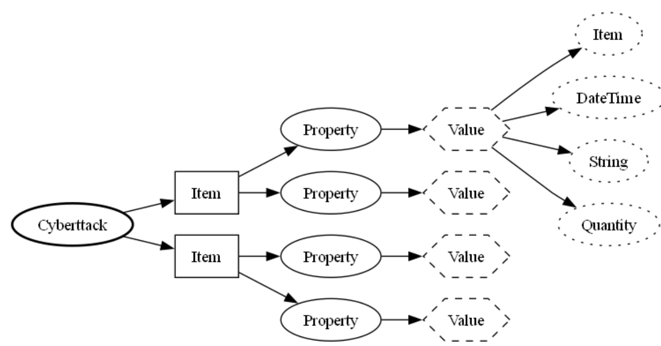

Fig. 3. An cyberattack, consisting of multiple items, each having multiple property-value pairs.

- An attack TTP is a class of all statements inferred during a forensic investigation of a cyberattack, i.e. all statements listed recursively from the item that represents an instance of a particular cyberattack.
- The cluster TTP of an AC is a union of all attack TTPs of cyberattacks already attributed to the AC, that is, all statements listed recursively from the item that represents an instance of an AC.

Next, we define:

- A match (or a matching statement) is when two statements with possibly different properties agree on their value, e.g., *"Malicious binary connects to an IP address XX.XX.XX.XX"* and *"A threat actor is known to use the IP address XX.XX.XX.XX"*. A strong match (or strongly matching statements) is a match when both values are identical between properties.
- A crossing between an attack TTP and a cluster TTP is a subset of their matching statements.

- An attribution hypothesis is a crossing that intersects attack TTPs only of ACs that are "said to be the same".
- An attribution is a attribution hypothesis that is concluded to be the most probable AC behind the attack.

Many crossings across multiple ACs may emerge when correlating broad or nonspecific statements. Even if a crossing intersects only related cluster TTPs, forming a attribution hypothesis, it doesn't necessarily indicate a true threat actor, as accuracy depends on data quality and potential false flags.

We may encounter another attribution hypothesis intersecting a different AC's cluster TTP. For example, if two distinct malware families — each tied to rival geopolitical ACs — are found on the same endpoint, at least three competing attribution hypotheses arise: actor A used both their tools and those of their rival as a false flag; actor B did the same; or both actors coincidentally attacked the same target. This illustrates that even a attribution hypothesis exclusive to one cluster TTP may not be conclusive. Ideally, all plausible hypotheses should be identified and shown to be less likely than the attribution claim.

The most common obstacle in the attribution process is incompleteness of data due to the following:

- The collection of the attack was limited, i.e., the forensic survey failed to acquire some IoCs.
- The forensic examination was insufficient; i.e., important connecting statements were not derived or even incorrectly interpreted.
- No relevant cluster in the threat landscape to which activity can be attributed (attribution of a new attack would fail because the AC behind it has not been observed yet).

After an attribution verdict, the question arises: should the attack TTPs be merged with those of the attributed AC? Even if the underlying crossing intersects only one cluster TTP, the link may be correlational rather than causal—the true actor might differ. This can result from incomplete data, overlooked stronger crossings, or attacker-inserted false flags. Blindly expanding cluster TTPs with each attributed attack can dilute specificity, creating clusters without a clear attribution hypothesis. Therefore, we recommend merging TTPs only after a confirmed attribution verdict. For attribution hypotheses, the new attack TTP should remain a separate AC, linked with the property *"said to be the same as"* - see Fig. 2 for illustration.

IV. Conclusion

This paper has explored the existing methodologies, frameworks, and challenges associated with cyber attribution, highlighting the trade-offs between exactness and verifiability in attribution efforts. Our proposed methodology aims to address these challenges by formalizing the attribution process and introducing structured approaches to evaluate the reliability of evidence. By integrating multiple data sources, refining attribution models, and ensuring a balance between exactness and verifiability, this work aims to enhance the robustness and overall quality of cyber threat attribution efforts.

We plan to utilize the groundwork laid in this paper to create a digital system for effective storage and correlation of data used as evidence in the attribution process. We are currently working on evaluating the best approaches for storing and working with data used for attribution.

Acknowledgement

This research was supported by CTU grant SGS23/208/OHK3/3T/18. The authors also acknowledge ESET Research for supporting the project.

References

[1] E. Hutchins, M. Cloppert, R. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," in Leading Issues in Information Warfare & Security Research, 2011.

[2] S. Caltagirone, A. Pendergast, Ch. Betz, "The Diamond Model of Intrusion Analysis," DOI:10.13140/RG.2.2.31143.56481, 2013.

[3] T. Rid, B. Buchanan, "Attributing Cyber Attacks," Journal of Strategic Studies (38). DOI: 10.1080/01402390.2014.977382, 2015.

[4] T. Steffens, "Attribution of Advanced Persistent Threats: How to Identify the Actors Behind Cyber-Espionage," DOI: 10.1007/978-3-662-61313-9, 2020.

[5] C. Boot, A. C. Serban, E. Poll, "Supervised Learning for State Sponsored Malware Attribution"

[6] J. Carr, "Faith-Based Attribution", medium.com, 2016.

[7] H. Berghel, "On the Problem of (Cyber) Attribution", 2017.

[8] F. Manola, E. Miller, B. McBride, "RDF primer 1.1: W3C Working Group Note 24 June 2014," 2014.