A Comprehensive Survey of Advanced Persistent Threat Attribution: Taxonomy, Methods, Challenges and Open Research Problems

Nanda Rani, Bikash Saha, Sandeep Kumar Shukla　*Fellow, IEEE,*

Abstract—**Advanced Persistent Threat (APT) attribution is a critical challenge in cybersecurity and implies the process of accurately identifying the perpetrators behind sophisticated cyber attacks. It can significantly enhance defense mechanisms and inform strategic responses. With the growing prominence of artificial intelligence (AI) and machine learning (ML) techniques, researchers are increasingly focused on developing automated solutions to link cyber threats to responsible actors, moving away from traditional manual methods. Previous literature on automated threat attribution lacks a systematic review of automated methods and relevant artifacts that can aid in the attribution process. To address these gaps and provide context on the current state of threat attribution, we present a comprehensive survey of automated APT attribution. The presented survey starts with understanding the dispersed artifacts and provides a comprehensive taxonomy of the artifacts that aid in attribution. We comprehensively review and present the classification of the available attribution datasets and current automated APT attribution methods. Further, we raise critical comments on current literature methods, discuss challenges in automated attribution, and direct toward open research problems. This survey reveals significant opportunities for future research in APT attribution to address current gaps and challenges. By identifying strengths and limitations in current practices, this survey provides a foundation for future research and development in automated, reliable, and actionable APT attribution methods.**

Index Terms—**Advanced Persistent Threat, Cyber Threat Attribution, Malware Attribution, Cyber Attack Attribution, Threat Profiling, Cybersecurity.**

I. Introduction

When firewalls and antivirus softwares become mere hurdles rather than barriers, we enter the realm of Advanced Persistent Threats (APTs). APT attacks are attack campaigns orchestrated by highly organized and often state-sponsored threat groups that operate covertly and methodically over prolonged periods [12]. APTs set themselves apart from conventional cyber-attacks by their stealthiness, persistence, and precision in targeting [1]. Unlike typical cyber-attacks, which aim for immediate and indiscriminate disruption or data exfiltration, APTs are orchestrated by resourceful adversaries who carefully conduct multi-phase operations. These adversaries excel at evading detection over prolonged periods, continuously adapting to security measures. APTs are recognized for

Nanda Rani, Bikash Saha, and Sandeep Kumar Shukla are with Department of Computer Science & Engineering, Indian Institute of Technology Kanpur, India (email: nandarani@cse.iitk.ac.in; bikash@cse.iitk.ac.in; sandeeps@cse.iitk.ac.in)

their underlying intent and the sophisticated attack methodology [1]. APTs employ meticulously crafted tactics, techniques, and procedures (TTPs) tailored to breach specific targets or organization, in contrast to traditional attacks that often rely on automated tools with widespread and indiscriminate targets. This high degree of customization and persistence makes APTs exceptionally difficult to detect and counteract.

The intricate and sustained nature of APTs demands a deep understanding of the actors orchestrating these campaigns, along with effective detection mechanisms. This deeper analysis is essential to anticipate future moves and tailor defenses accordingly. The process of identifying and linking sophisticated attacks back to the associated and responsible threat actors or nation-states is known as APT Attribution [7], [24], [43], [88]. It is vital for understanding the complex nature of these threats and the actors behind them, providing crucial insights into the broader context of cyber activities on a global scale. Threat attribution remains a cornerstone in cybersecurity, as it helps to reveal the intricate web of motives and strategies employed by these malicious entities.

Attributing APT attacks is instrumental in understanding the attacker's motivations, strategies, and techniques. Further, such understanding, in turn, informs the development of targeted defensive strategy, proactive countermeasures and enhances the resilience of potential targets [123]. Beyond bolstering cybersecurity, precise attribution is the groundwork for legal and diplomatic actions. It enables victims to seek justice through legal channels, impose sanctions, or engage in diplomatic efforts to curb malicious activities [7]. In recent times, the FBI (Federal Bureau of Investigation) charged Lazarus group [86] linked three individuals: Park Jin Hyok [99], Kim Il [122], and Jon Chang Hyok [134]. On a broader scale, accurate and timely attribution may act as a deterrent. It also signals to potential aggressors that their actions will not remain anonymous, and they will be held accountable. Further, attribution facilitates the sharing of actionable threat intelligence across organizations and sectors, which further improves the overall cybersecurity posture of the organization [109]. Thus, threat attribution is crucial in immediate response and long-term security planning.

Attributing attacks involves a multifaceted approach, combining digital forensics, intelligence analysis, and geopolitical insights. These approaches peel back the layers of anonymity or evasiveness that attackers hide behind [36], [61]. This intricate process employs a blend of technical and non-technical artifacts to identify the unique digital fingerprints

TABLE I: Comparison with related existing survey

Paper	Year	Attribution Case Study	Artifact Taxonomy	Method Classification	Dataset Classification	Challenges	Open Research Problems
[88]	2022	✗	✗	✓	✗	✗	✗
[98]	2022	✗	✗	✓	✗	✓	✗
[43]	2023	✓	✗	✗	✗	✗	✗
[42] *	2024	✗	✗	✓	✓	✓	✓
Our Survey	2024	✓	✓	✓	✓	✓	✓

* This survey is specific to malware-based attribution methods

left by attackers. Threat analyst teams scrutinize these indicators, comparing them against vast databases of known threats. It further aids in finding similarities or establishing new profiles for unidentified attackers. Analysts perform this analysis on traditional frameworks such as Diamond Model [3], CAM [24], Q model [17], triangle model [6], and MICTIC framework [7], which facilitate multi-dimension approach to connect the dots between collected artifacts and suspected threat actor. While traditional manual methods allow for deep and contextual analysis of attacks, their effectiveness and scalability are increasingly challenged by the rapidly evolving cyber threat landscape [17], [88]. The advent of advanced automated technologies such as artificial intelligence (AI) and machine learning (ML) presents an opportunity to enhance APT attribution's performance, efficiency, reliability, and scalability. Further, the literature of automated threat attribution leverages series of attack artifacts with state-of-the-art automated technological capabilities to uncover the threat actors. We review such available automated threat attribution methods and present a comprehensive survey of automated APT attribution.

A. Motivation

The motivation behind this survey on automated APT attribution stems from the growing frequency and sophistication of APT attacks, which necessitates scalable and efficient attribution methods. Advanced technologies such as AI and ML offer opportunities to enhance the performance, reliability, and scalability of these methods. Despite the availability of several automated approaches in the literature, they are dispersed and lack a systematic synthesis. This survey aims to consolidate and critically analyze these methods, highlighting strengths, potential limitations and understanding their real-world applications. The goal is to advance the field by identifying effective strategies for automated APT attribution and addressing remaining gaps and challenges.

B. Review of prior surveys and gap analysis

The current surveys and research papers have explored various aspects of APTs and attribution techniques. Previous surveys have predominantly focused on traditional attribution methods [43], [88], the detection and mitigation of APTs [1], [43], [56], profiling techniques [98], threat intelligence of APTs [2], [9], [62], and threat modelling of APTs [20]. While these surveys provide valuable insights about APTs, there is a notable gap in the literature concerning the comprehensive review of automated APT attribution methods. The literature

on the APT attribution methods is dispersed and scattered. In our exploration, we find very few surveys present in the literature [42], [43], [88], [98], which discuss threat attribution, but that too lacks detailed discussion, comparison of available attribution methods and current attribution challenges. There is also a lack of systematic literature that can summarize the complete flow and end-to-end automated attribution process. Such summarization facilitates a broader understanding of the community. Therefore, this paper systematizes the automated APT attribution methods available in the literature. The comprehensive survey aids in understanding the available effective automated attribution methodology. Table I compares this survey with the related existing study.

C. Scope and contribution

This survey consolidates current knowledge, highlights recent advancements, and identifies gaps in existing methodologies. It also serves as a bridge between various isolated studies and offers a unified perspective that is currently lacking. This consolidation is crucial for advancing research in this domain and provides cybersecurity practitioners and policymakers with a comprehensive overview of effective methods. This survey thoroughly reviews the state-of-the-art techniques for automated APT attribution by presenting a detailed case study and assessing the effectiveness of various automated methods. The contributions of this paper include review of available attribution dataset and methods, a critical analysis of current methodologies, identification of existing gaps and challenges, and presenting the open research problem toward effective and reliable automated APT attribution. This survey thus serves as a valuable resource for researchers, blue and red teams of the organizations and practitioners in cybersecurity. It also offers insights and guidance for advancing the practice of automated APT attribution. The key contributions of this survey are the following:

(a) We contextualize the evolution of APT attribution by classifying the advancements in the attribution methods since beginning.

(b) We taxonomize the dispersed attribution artifacts, which can be utilized effectively in the attribution process.

(c) We systematically illustrate the end-to-end attribution process, detailing each phase from data collection to analysis.

(d) We explore and classify all (structured and unstructured) publicly available threat data, which can be leveraged to develop a dataset for presenting new attribution methodologies.

(e) We classify the current literature on attribution methodologies and comprehensively compare these methodologies based on various bases.

(f) We examine the current technical and methodological challenges in automated attribution that hinder accurate and reliable threat attribution.

(g) We identify and propose potential open research problems in automated attribution, suggesting areas where further investigation and innovation can significantly advance threat attribution.

D. Structure of the Paper

The structure of this paper is shown in Fig 1. Section II presents the current state of APT attribution, and Section III details an attribution case study of past attack campaigns. Then, the proposed attribution artifact taxonomy is presented in Section IV, and a survey of the current attribution datasets and methods is discussed in Section V. Further, we present the current attribution challenges in Section VI and discuss open research problems and future research directions in Section VII. Section VIII concludes the presented survey. The list of acronyms used in this paper is listed in Table II.

TABLE II: List of acronyms and their corresponding descriptions

Acronym	Description
AI	Artificial Intelligence
API	Application Programming Interface
APT	Advanced Persistent Threat
ATT&CK	Adversarial Tactics, Techniques and Common Knowledge
C&C	Command and Control
CFG	Control Flow Graph
CTI	Cyber Threat Intelligence
CVE	Common Vulnerabilities and Exposure
DNN	Deep Neural Network
DNS	Domain Name System
DT	Decision Tree
GNN	Graph Neural Network
IoC	Indicators of Compromise
KNN	K-nearest Neighbor
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning
NB	Naïve Bayes
OSINT	Open Source INTelligence
RF	Random Forest
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TTP	Tactics, Techniques, and Procedures
XGBoost	Extreme Gradient Boosting

II. APT ATTRIBUTION

Attribution in the context of cybersecurity, particularly concerning APTs, is distinct from traditional cyber criminal investigations. Attribution involves tracing the source of cyber attacks, typically tied to state-sponsored entities. The evolution of cyber threat attribution methodologies has seen significant advancements over the decades.

In the early beginnings (1986-2000), manual methods prevail, exemplified by the "Cuckoo's Egg" incident [49], [88], where digital forensics plays a crucial role in tracking cyber intrusions. During this period, the U.S. military also introduces the concept of attribution [88], laying the groundwork for
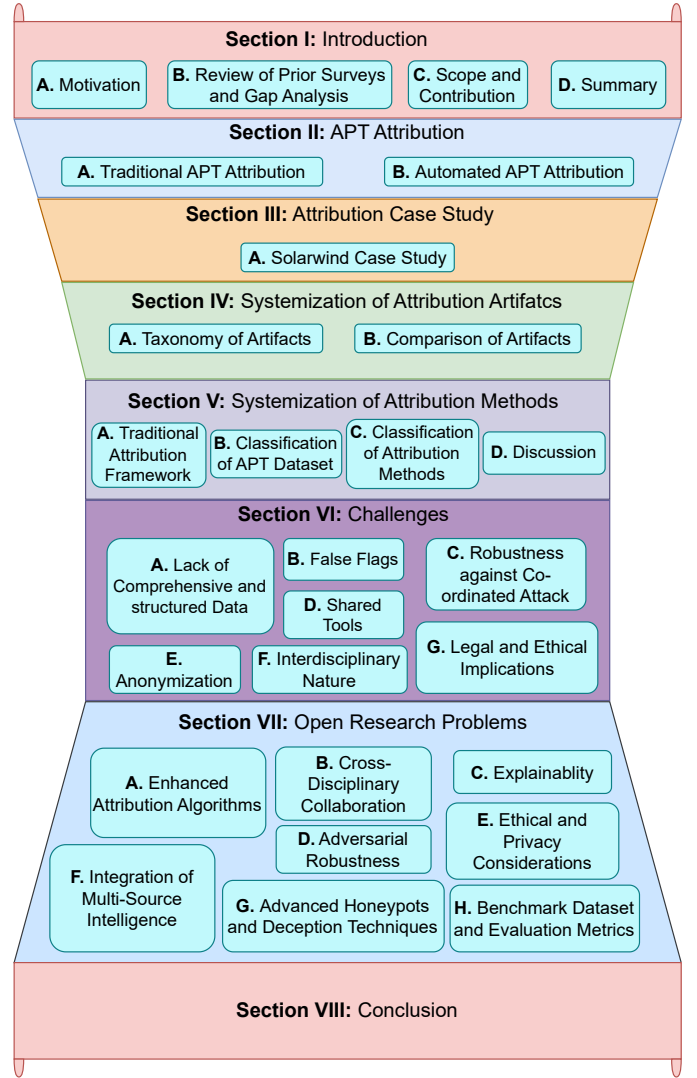

Fig. 1: Organization of this paper

systematic approaches. The goal was to identify the actors behind cyberattacks, particularly nation-states and organized cybercriminal groups. Initial attribution efforts were hindered by the lack of standardized methodologies and the complexity of APT attack vectors. As cyber incidents grew more complex and prevalent, the need for structured approaches became apparent, leading to the development of foundational frameworks and methodologies.

From 2000 to 2015, a more systematic approach emerges to enhance coordination and the development of frameworks such as the Diamond Model [3] and MITRE ATT&CK [118], which provide structured methods for documenting adversary tactics, techniques, and procedures. The documented adversary's modus operandi are correlated to identify any link between attack incidents with historical attacks.

With the emergence of automation and ML in analyzing vast datasets quickly and efficiently, the period from 2015 to 2023 witnesses the rise of automated approaches along with behavioral analytics and enhanced threat intelligence sharing to improve attribution accuracy.
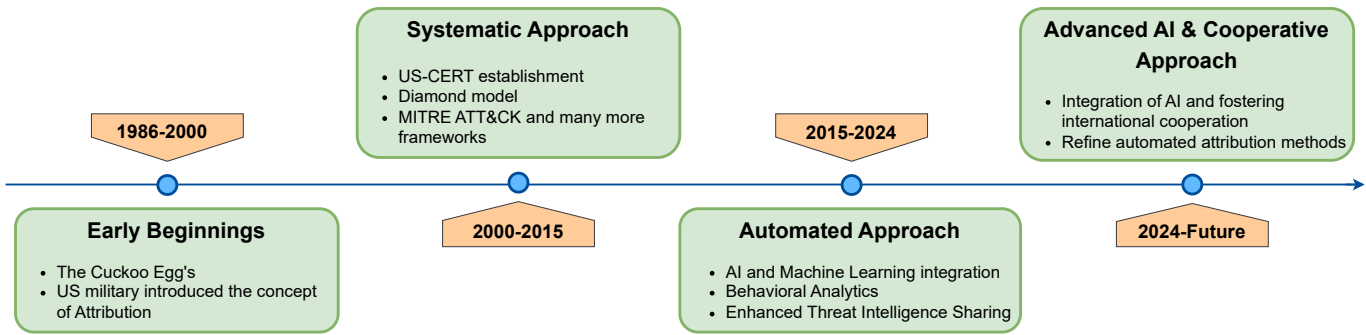
Fig. 2: Evolution of cyber threat attribution

Looking forward, the advanced AI and cooperative approach (2024-future) emphasize the continued integration of AI and fostering international cooperation, aiming to refine automated attribution methods and enhance their robustness against evolving cyber threats [105]. A timeline of evolution of attribution methodologies is shown in Fig. 2.

A. Traditional APT attribution

With the evolution of cyber threats, the techniques for attribution have also advanced towards structure-based frameworks. In the period of 2000 to 2015, the cybersecurity community or researchers shift their focus to develop more systematic approaches. During this period, we observed the rise of frameworks such as the Lockheed Martin Cyber Kill Chain (CKC) [82], MITRE ATT&CK framework [118], Diamond Model [3] and MICTIC framework [7]. The CKC provided a structured methodology for analyzing and responding to cyber threats. Concurrently, the MITRE ATT&CK framework [118] emerged, offering a comprehensive knowledgebase of TTPs used by threat actors. Further, the attribution framework such as Diamond Model [3] and MICTIC framework [7] presents a structured way to categorize attribution artifacts on a different basis and employ expert reasoning to compare with historical databases. These frameworks enabled a more granular analysis of cyber incidents and facilitated the identification of patterns and signatures associated with specific threat actors. The progression of these methodologies marked a significant leap in the ability to not only detect but also attribute cyberattacks with greater precision [88].

Traditional methods rely heavily on the expertise of cybersecurity analysts and investigators. Security analysts meticulously examine digital forensic evidence, such as malware signatures, IP addresses, and domain names used during the attack period. This process often involved correlating technical indicators with known threat actor profiles, analyzing linguistic and behavioral patterns, and sometimes leveraging intelligence from law enforcement and intelligence agencies [7], [25], [64], [76], [77]. The traditional methods provide deep and contextual attribution information but face several limitations. The primary limitation of framework-based attribution lies in its inherent limitations regarding speed and scalability. In an era of rapidly increasing cyber-attacks, the time-consuming nature of manual analysis becomes a significant bottleneck. The delay in attribution hampers the timely response to ongoing threats and affects the ability to deter future attacks. Furthermore, the reliance on expert analysis is invaluable for understanding complex cyber espionage campaigns or sophisticated state-sponsored attacks, but introduces subjectivity, potential bias, and scalability issues in the attribution process.

B. Automated APT attribution

Traditional attribution's complexity and resource demands have led to the development and adoption of automated APT attribution methods. Automated attribution leverages advanced technologies, including ML, AI, and big data analytics, to enhance the speed and accuracy of identifying threat actors. By automating the analysis of vast amounts of data from diverse sources, such as network logs, threat intelligence feeds, and malware repositories, these systems can detect patterns and correlations that human analysts might miss. The benefits of automated APT attribution are manifold:

(a) It significantly reduces the time required to identify and attribute attacks, enabling faster response and mitigation efforts.
(b) Automated systems can continuously learn and adapt, improving their performance over time as they process more data and encounter new threat patterns.
(c) Automation can help overcome some of the limitations of manual methods, such as the ability to analyze large-scale and complex datasets that would be impractical for human analysts to handle.
(d) Automated methods reduce the likelihood of human error and bias in analysis, ensuring a more consistent and objective approach to attribution.

As automated methods continue to advance, these methods are becoming indispensable for organizations aiming to identify threat actors and mitigate potential damage efficiently and quickly. Overall, the automated APT attribution represents a critical advancement in cybersecurity and offers a scalable and efficient approach to understanding and combating sophisticated cyber threats. In this paper, we used the word attribution and automated attribution interchangeably.
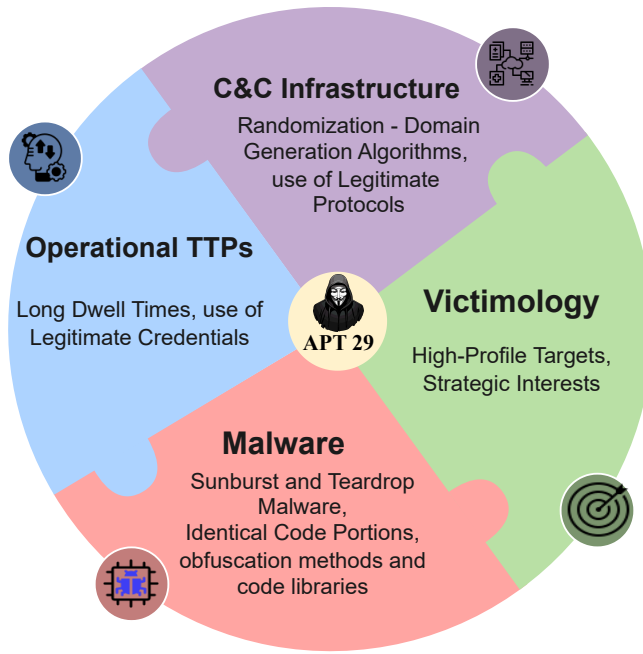
Fig. 3: Solarwinds Case Study: Attribution Artifacts

III. ATTRIBUTION CASE STUDY

To demonstrate the APT attack investigation and artifacts aid in attribution, we discuss a high-profile target campaign performed in 2020 employing supply chain-attack. This case study discusses the steps involved in tracing the attack, the techniques used by the threat actors, and the artifacts collected to accurately attribute the attack. By dissecting this high-profile incident, we aim to provide a comprehensive understanding of the methodologies and key artifacts which led to attribute this campaign.

A. SolarWinds cyberattack

In December 2020, the leading cybersecurity firm Fire-Eye (now Mandiant, a Google cloud company) discovered a widespread cyber espionage campaign. This attack targeted the software supply chain of SolarWinds[1] by compromising its Orion software to distribute malware to several high-profile organizations, including U.S. government agencies and private sector companies [129]. Several leading security firms collaborated to share intelligence and uncover the responsible nation-states for such a large-scale cyber attack [87]. The artifacts used by them to attribute the attack are shown in Fig 3 and described below.

(a) **Malware & Code Analysis:** The key artifact in the this attack was the Sunburst malware embedded within the Orion software updates [125]. After installation, it deploys secondary payloads like Teardrop malware, which enabled lateral movement within networks.

(b) **C&C Infrastructure:** The attackers employed a variety of domains for C&C purposes such as advanced evasion

---

[1]A major IT management company

---

techniques like randomization using DGA (Domain Generation Algorithm) and encryption, making detection and analysis challenging [60].

(c) **Operational TTPs:** The attackers' operational tactics also played a vital role in the attribution process. In the Solarwinds attack, the attackers demonstrated these advanced tactics by maintaining a low profile within the networks for several months, exfiltrating data without triggering alarms [23].

(d) **Victimology:** The selection of targets provided additional context for attribution. The attackers focused on high-profile entities, including U.S. government agencies, critical infrastructure, and major corporations [129].

The SolarWinds campaign showcases the complexity and sophistication of advanced cyber espionage campaigns. By examining artifacts such as malware analysis, C&C infrastructure, code overlaps, operational tactics, and victimology, cybersecurity experts attributed the campaign to APT29 (also known as Cozy Bear or Nobelium) [44], [129]. The reverse engineering revealed sophisticated coding, obfuscation techniques, and code similarities between the SolarWinds malware and previous APT29 campaigns [60], [125], providing key evidence for attribution. The use of DGA, randomization techniques, and the long dwell time were consistent with APT29's past behavior [44]. The targeted intelligence-gathering aligned with APT29's strategic objectives, further supporting the attribution of the SolarWinds espionage campaign to the Russia-backed group APT29 [113], [136]. The convergence of these multiple evidence provided a robust foundation for attribution. This analysis highlights the importance of comprehensive and multi-layered analysis in identifying and understanding sophisticated threat actors.

IV. SYSTEMIZATION OF ATTRIBUTION ARTIFACTS

Attribution is a multifaceted approach and is decided based on multiple clues. We call these clues as attribution artifacts. We define **attribution artifacts** as *any critical pieces of evidence or indicators or information which aid in attributing the cyber threats, i.e., linking cyber threats to specific threat actors or origins or sources.* during Russia's invasion of Ukraine, the AcidRain wiper malware was employed in the Viasat attack [89], [101], and further analysis uncovered code overlaps between AcidRain wiper and the SolarWinds malware [44].

Artifacts of threat attribution hold significant importance because they serve as the foundational elements that identify and link cyber threats to specific actor or group. These artifacts, including malware signatures and attack behavior patterns, are crucial in linking an incident to its perpetrators. By examining these pieces of evidence, analysts understand the source and intent of an attack, which can directly influence how an organization or government responds to and mitigates potential threats.

A. Attribution artifacts taxonomy

Artifacts plays a crucial role in the attribution process. It is essential to understand the various types that can help security

analysts link or profile the threat actor. Such understanding enhances defenders' ability to trace attack origins and identify the techniques used. In this section, we present the taxonomy of artifacts as shown in Fig. 4 Our taxonomy of attribution artifacts categorizes the diverse information that can be utilized in the cyber threat attribution process. Based on source of the attribution artifacts, we categorize them in two types: 1) Attack artifacts 2) Non-attack artifacts. The details about these artifact categories and further classification are discussed in this section.

1) Attack artifacts: It consists the artifacts which analysts captured from attacked infrastructure such as malware, IoCs, and toolchains etc. Based on nature of the artifacts, we categorize attack artifacts into two category: 1) Evidentiary artifacts and 2) Behavioral artifacts. The details about these artifact categories and further classification are discussed below.

a) **Evidentiary artifacts:** The tangible data or information providing direct evidence of a threat actor are evidentiary artifacts. These artifacts are crucial for identifying and analyzing specific attack elements as they can be directly observed and recorded. This includes direct digital evidence, such as IP addresses, cryptographic hashes, and email addresses, obtained from cyber incident investigations. Evidentiary artifacts offer concrete evidence that can link cyber activities to specific devices, networks, or individuals. We further divide evidentiary artifact into three types as explained below:

   i) *Indicators of Compromise (IoCs):* Specific pieces of evidence indicating a system has been compromised are IoCs. A direct link to specific threat actors or campaigns can be made if same IoCs are captured in two attacks. Concrete examples of evidentiary artifacts may include IP addresses, URLs, email addresses, cryptographic hashes, domain names, digital signatures, and certificates.

   ii) *C&C Infrastructure:* It includes the servers and domains that control and manage deployed malwares. Unique characteristics of C&C infrastructure can reveal the operational patterns of attackers, often tied to specific groups. This information details the resources attackers use to maintain communication with compromised systems—for example, hosting providers, communication protocols, and encryption methods.

   iii) *Cryptocurrency Transactions:* Transactions involving cryptocurrencies are often linked to ransom payments or the purchase of illicit services on the dark web. Blockchain analysis can sometimes trace these transactions to specific wallets and aid in connecting them to known groups or activities [108], [112], [126]. For example, analyzing wallet addresses and transaction patterns links transactions to known wallets.

b) **Behavioral Artifacts:** The methods attackers use, including their operational patterns and decision-making processes, are known as behavioral artifacts. Understanding these behaviors can reveal the attackers' level of sophistication, potential motives, and operational goals. It is crucial to identify the likely perpetrators behind cyber attacks. We further divide behavioral artifact into five types as explained below:

   i) *Tactics Techniques and Procedures (TTPs):* It describes how threat actors operate and offers valuable insights into their behavior and capabilities. Specific TTPs can act as signatures for certain groups, revealing their unique modus operandi such as using phishing to perform initial access or employing PowerShell for execution. The TTPs can be extracted from historical threat reports [81], [90], [100] or threat intelligence feeds to create the dataset. The TTP extraction methods help attackers to achieve their objectives and provide a detailed picture of their operational approach. The MITRE ATT&CK framework includes 12 different type of tactics as shown in Fig 4.

   ii) *Malware & Code analysis:* The artifacts related to malware analysis and the source code of malware are crucial for understanding its functionality, propagation methods, and origins. Unique code snippets, code reuse, or targeted vulnerabilities can link malware to known threat actors [18]. For example, unique malware functions and reusing specific code blocks seen in other known malware can provide valuable insights. The capabilities of the malwares can also be extracted and correlate them to identify similar ones [124], [137]. The process of dissecting malware helps identify its behavior and origin. It involves performing static, dynamic, and hybrid analysis of malware samples.

   iii) *Toolchains:* It includes the tools utilized by attackers throughout the lifecycle of an attack. The preference for certain tools or specific combinations can indicate particular threat actors or communities [42]. The custom-built hacking tools and well-known exploitation frameworks can provide clues about the attackers' identities and affiliations. These tools and software are essential for attackers to conduct the operations such as exploit kits, remote access tools, and custom scripts.

   iv) *Language and Writing Styles:* Linguistic analysis can indicate the native language of the attackers and potentially their country of origin [7]. For example, using specific dialects or idioms in phishing emails and error messages in malware can provide valuable insights. These linguistic characteristics help identify and profile the attackers. It includes code comments, phishing email text, and ransom notes.

2) Non-attack Artifacts: It consists the artifacts which analysts captured from outside of the attacked infrastructure such as threat intelligence feeds, public claims, and geopolitical tension etc. Based on nature of the artifacts, we categorize non-attack artifacts into two category: 1) Open Source INTelligence (OSINT) artifacts and 2) External artifacts. The details about these artifact categories and further classification are discussed below.

a) **Open Source INTelligence (OSINT) Artifacts:** It includes the artifacts derived from publicly available sources, such as social media, threat reports, technical databases of historical attacks, and hacker forums. These
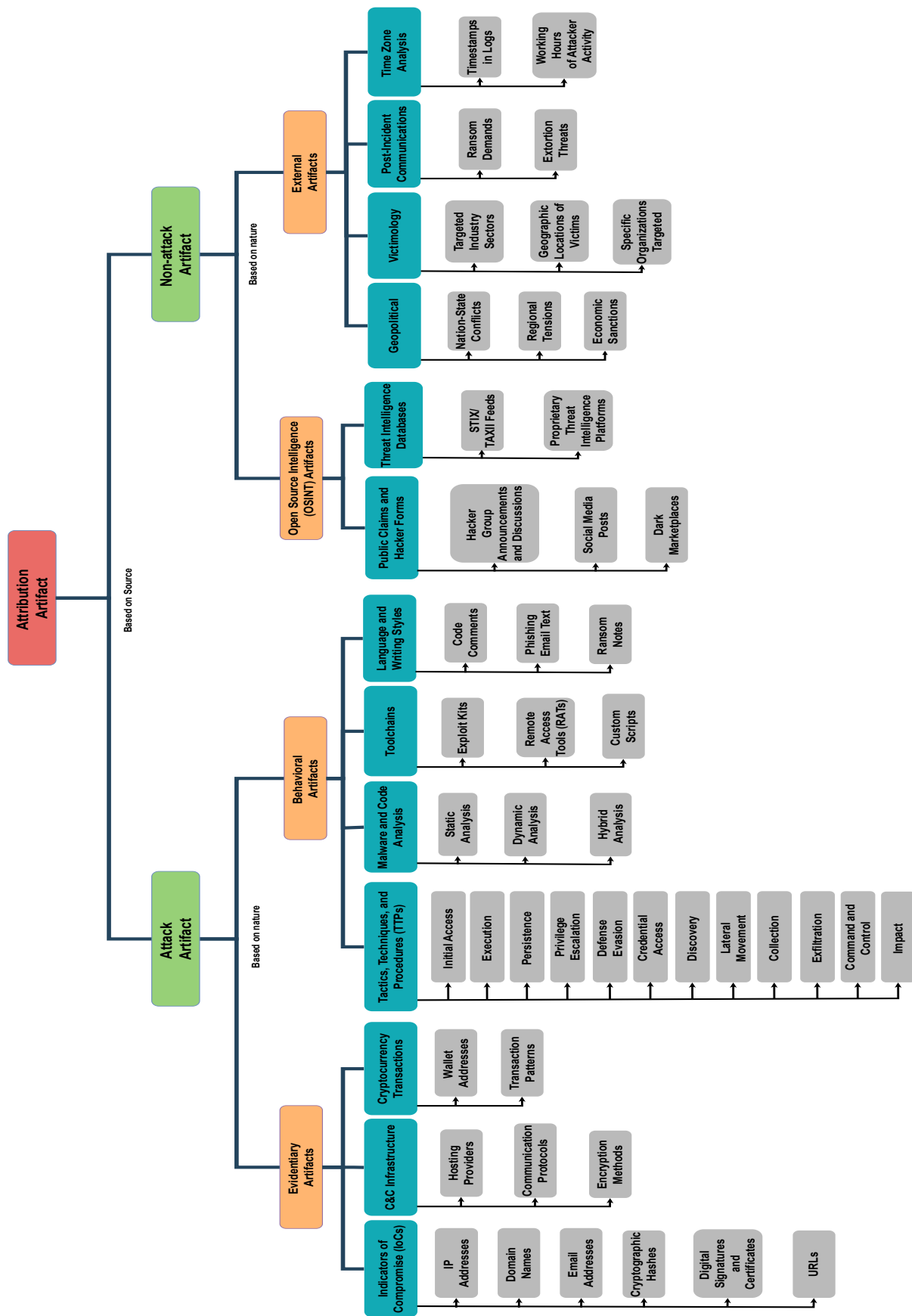
Fig. 4: Taxonomy of attribution artifacts

artifacts can provide contextual and corroborative information to support attribution by linking public disclosures or activities to specific cyber incidents or threat actors. OSINT artifacts offer valuable insights into the threat actor's activities, intentions, and associations. We further divide this artifact into three types, explained below:

i) *Public Claims and Hacker Forms:* The online platforms, including hacker forums, dark websites, and social media channels, are crucial sources of information where cybercriminals discuss tactics, share tools, and claim responsibility for cyber incidents. Although these claims need verification, they can provide direct links to the perpetrators. For instance, threat group blogs on dark websites, telegram channels, and social media posts by groups claiming an attack are crucial sources of information [95], [132]. Information or tools shared on public platforms can often be directly linked to specific cyber incidents or groups, such as a thread in a dark web forum where an actor sells access credentials to compromised systems.

ii) *Threat Intelligence Databases:* The databases aggregate and analyze IoCs, attack patterns, and threat actor profiles. By correlating data from historical threat databases with indicators captured during an incident, analysts can establish links and identify related known threat actors. These repositories of threat data are invaluable for identifying and understanding cyber threats. For example, STIX/TAXII feeds and proprietary threat intelligence platforms.

b) **External Artifacts:** It includes information about the broader context in which cyber threats occur such as geopolitical events and international cyber policies. These artifacts provide insights into the external motivations and constraints affecting cyber threat actors. By analyzing them, we can potentially identify why an attack was carried out and who might be responsible based on alignment with geopolitical motives. We categorize these artifacts into four types:

i) *Geopolitical:* Information related to international relations, conflicts, and political events that may motivate or be exploited by cyber threat actors is crucial. We can identify possible state sponsorship or politically motivated actors by correlating cyber attacks with geopolitical events [7]. A rise in cyber espionage targeting government agencies during periods of heightened diplomatic tensions between two countries can indicate geopolitical motives. We witnessed such campaigns during recent Russia-Ukraine conflicts [37], [57].

ii) *Victimology:* The study of victims targeted by cyberattacks helps to identify patterns that can reveal the attackers' motives [65], [110]. Analyzing victim selection provides insights into the attackers' objectives and potential identities, allowing us to link them to specific groups or goals. This has been thoroughly discussed in CAM (Cyber Attribution Model) [77]. Key areas to examine include targeted industry sectors, the geographic locations of victims, and specific organizations that were attacked.

iii) *Post-Incident Communication:* Communication after an incident can provide valuable insights into the attackers' motives or negotiation strategies [96]. For instance, ransom notes, negotiation attempts, or taunting messages can reveal important details about the attackers' intentions and tactics. These communications often occur directly between the attacker and the victim, offering a glimpse into the dynamics of the attack. Examples include ransom demands and extortion threats.

iv) *Time Zone Analysis:* Analyzing the time zones in which attackers operate can provide clues about their origins and working patterns [7]. By examining timestamps in logs and the hours during which attacker activity is most frequent, we can infer the likely geographic locations of the threat actors. This information helps narrow down potential suspects and understand the operational habits of the attackers. Examples include evaluating timestamps in system logs and identifying the working hours of attacker activity.

B. Comparison of automated attribution artifacts

For accurate threat attribution, it is crucial to analyze and compare the effectiveness of various artifacts [24], [45]. Different artifacts, such as malware signatures, network traffic logs, and OSINT data, provide unique insights into an attack. Security analysts can discern patterns and link behaviors to specific threat actors by systematically evaluating these artifacts. Therefore, we compare the classified artifacts (as shown in Fig. 4) based on the different metrics: Relevance, Integrity, Credibility, Timeliness, and Accessibility. We compare artifacts based on these metrics with respect to three scale points: Low (○), Medium (◐), and High (●). These scale points represent confidence in corresponding measures, i.e., low means minimal confidence, medium represents moderate confidence, and high means more confidence. Table III shows the comparison of all classified artifacts.

Relevance measures how directly an artifact contributes to the identification of a specific threat actor. High relevance artifacts, such as IOCs, TTPs, C&C infrastructure, and malware & code analysis, provide strong and often unique identifiers that can clearly tie activities to specific threat actors, making them crucial in the attribution process. Medium relevance artifacts, like toolchains, public claims, threat intelligence database, geopolitical, and time zone analysis, offer valuable context but may require additional corroboration to be definitive, reflecting their role in supporting rather than leading attribution efforts. Low relevance artifacts, such as cryptocurrency transactions, language and writing styles, victimology, and post incident communication, provide useful supplementary information but often lack the direct connection needed for strong attribution, making them less central to the process. This analysis aligns with the idea that artifacts vary in their directness and specificity when attributing cyber threats.

Integrity assesses the resistance of an artifact to manipulation or deception. High integrity artifacts like TTPs, victimology, and malware & code analysis are reliable because they are

TABLE III: Comparison of Attribution Artifacts

Source based	Nature based	Artifact Name	Relevance	Integrity	Credibility	Timeliness	Accessibility
Attack Artifacts	Evidentiary Artifacts	Indicator of Compromise (IoCs)	●	○	●	●	●
		C&C Infrastructure	●	○	●	◐	◐
		Crypto Currency Transaction	○	○	●	○	○
	Behavioral Artifacts	Tactics, Techniques & Procedures (TTPs)	●	●	●	◐	◐
		Malware & Code Analysis	●	●	●	◐	○
		Toolchains	◐	○	●	◐	◐
		Language & Writing Style	○	◐	●	○	◐
Non-attack Artifacts	OSINT Artifacts	Public Claims and Hacker Forums	◐	○	○	◐	●
		Threat Intelligence Databases	◐	◐	◐	○	●
	External Artifacts	Geopolitical	◐	◐	◐	○	◐
		Victimology	●	●	●	◐	◐
		Post-Incident Communication	○	○	●	○	◐
		Time Zone Analysis	◐	○	●	●	●

challenging for attackers to fake or alter without compromising their operational capabilities, making them robust indicators for attribution. Medium integrity artifacts, such as language & writing style, threat intelligence databases and geopolitical analysis, provide useful information but can be subject to interpretation, bias, or imitation, requiring careful validation. Low integrity artifacts, like IOCs, C&C infrastructure, toolchain and public claims, are more easily manipulated or spoofed, reducing their reliability in making confident attributions. This analysis underscores the importance of using artifacts with high integrity to ensure accurate and reliable cyber threat attribution.

Credibility refers to the reliability of the source from which an artifact is obtained. It measures how likely it is that the information provided by the source is accurate and unbiased. High credible artifacts, such as those providing TTPs, C&C infrastructure data, TTPs, toolchains, and malware analysis, are typically derived from detailed investigations by reputable cybersecurity professionals and organizations, ensuring the reliability and accuracy of the information. Medium credible artifacts, such as geopolitical analysis, and threat intelligence dataases, offer useful insights but may come from a mix of reliable and less verified sources, requiring cross-verification to ensure accuracy. Low credible artifacts, such as public claims and hacker forums, are often unverified and can be intentionally misleading, making them less trustworthy for confident attribution. This analysis emphasizes the need to prioritize artifacts from high-veracity sources to ensure accurate and credible cyber threat attribution.

Timeliness measures how quickly an artifact can be obtained and analyzed after an attack. High timeliness artifacts, such as IOCs, toolchains, and time zone analysis, are often available almost immediately during or after an attack, allowing for rapid response and mitigation. Medium timeliness artifacts, like C&C infrastructure, TTPs, victimology, public claims, and malware & code analysis typically require some time for monitoring, analysis, and pattern recognition, which may delay immediate action but still offer relatively prompt insights. Low timeliness artifacts, such as cryptocurrency transactions, language & writing styles, threat intelligence databases, post incident communications and geopolitical considerations, involve more extensive investigation and slower information gathering,

making them less suited for real-time threat response. This analysis emphasizes the importance of prioritizing timely artifacts in situations where rapid attribution and response are critical.

Accessibility refers to how easily an artifact can be collected and analyzed during the attribution process. High accessibility artifacts like IOCs, public claims, threat intelligence databases and time zone analysis are readily available and require minimal effort to collect and analyze, making them easier for defenders to leverage in attribution. Medium accessibility artifacts, such as TTPs, C&C infrastructure, toolchains, language & writing analysis, geopolitical, victimology, and post incident communications involve a moderate level of complexity and require some specialized knowledge and tools, reflecting the need for more effort to effectively utilize them. Low accessibility artifacts, such as malware & code analysis and cryptocurrency transactions, demand advanced technical skills and resources, akin to the top of the Pyramid of Pain, where the effort to analyze and attribute becomes more challenging and resource-intensive. This analysis highlights how the accessibility of artifacts aligns with the effort required to disrupt or attribute adversarial activities.

This comparative analysis enhances the reliability of attribution, enabling more precise identification of adversaries. A glimpse of the reasoning behind this analysis is provided in Table VIII in the Appendix. Moreover, understanding the strengths and limitations identified during comparison of each artifact type ensures a comprehensive approach, reduces the likelihood of false attributions, and improves overall threat intelligence.

V. SYSTEMIZATION OF AUTOMATED ATTRIBUTION METHODS

In this section, we systemize automated attribution methods, detailing each aspect comprehensively. First, we outline the framework followed for automated attribution. Next, we discuss publicly available datasets that facilitate the analysis and development of attribution methods. We then delve into the current literature on automated attribution methodologies, providing a detailed categorization of datasets and methods. Finally, we discuss critical comments and comparisons

of existing automated attribution methods, mentioning their strengths and limitations.

A. Automated Attribution Framework

APT attribution leverages advanced technology to automate the process of identifying the actors behind APT campaigns without or with minimal human intervention. The process leverages AI, ML, and big data analytics to analyze the TTPs and other artifacts used in cyber attacks. The analysis includes comparison or correlation of given artifacts against extensive databases of known threat actors and their historical activities. The goal is to rapidly and accurately identify the perpetrators by understanding their motives, methodologies, and potentially their affiliation. The typical automated APT attribution process involves four key steps as shown in Fig 5 and described below:

(a) **Data Collection:** The first step includes gathering the vast amounts of data from various sources, including network traffic, logs, malware samples, and threat intelligence feeds. These data provide the raw material for analysis.

(b) **Preprocessing and Enrichment:** The collected data is cleaned, normalized, and enriched to facilitate analysis. This step may involve extracting features from raw data such as malware samples, network and system logs, correlating related data points, and filtering out irrelevant information.

(c) **Training or Pattern Recognition:** Using ML algorithms, the system analyzes the preprocessed data to identify patterns and behaviors that match known APT groups. This analysis can include the examination of malware signatures, communication patterns, and attack methodologies.

(d) **Attribution:** Once analysis or training is performed, the system attributes the attack to specific APT groups by matching identified patterns with known threat actors. This step relies on a comprehensive database of threat actor profiles, including their TTPs, preferred malware, targets, and historical attacks.
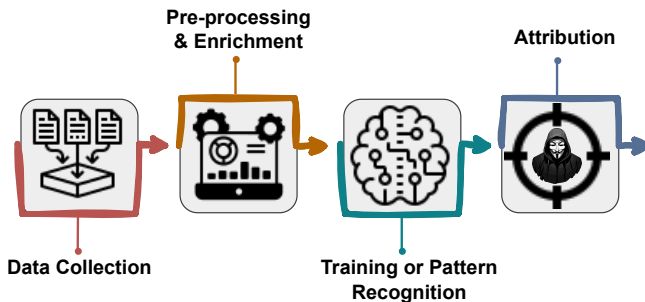
Fig. 5: Automated APT Attribution Framework

The automated attribution process begins with extensively curating APT data from both public and private repositories. Next, preprocessing and feature engineering are conducted. The engineered features are then utilized to learn the attribution aspects, culminating in the final attribution process.

B. APT Attribution Dataset Classification

We curate the publicly available datasets, which can be used to develop APT attribution methodology. In our exploration, we classify public data sources based on the type of data they consist. Therefore, we classify the data repositories into four types: Malware dataset, Threat report dataset, Attack pattern dataset, and Heterogeneous dataset. The distribution is shown in Fig. 6. We find a total of twelve public sources that facilitate data collection specific to APT attacks, which we discuss in detail below:
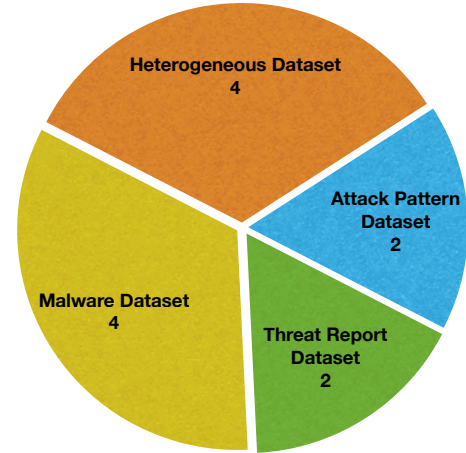
Fig. 6: Dataset Classification of APT Attribution

1) Malware Dataset: Researchers extensively utilize malware associated with APT groups from past attack campaigns to develop attribution methodologies, such as [14], [22], [29], [53], [72], [79], [85], [94], [116], [131]. To develop the methodology, several researchers have gathered malware samples from various repositories and labeled their corresponding APT groups using OSINT databases. We name such collection of APT malware along with their associated threat group dataset as malware dataset for attribution. We discuss all such public dataset collections in this section.

The authors in [31], [39] compiled an diverse and balanced APT malware dataset using malware samples and their corresponding sandbox reports. They obtained details on state-sponsored malware, like sample hashes and corresponding threat group name, from threat intelligence reports published by security firms like Trellix [73], F-Secure [93], and Kaspersky [120]. They leveraged VirusTotal [16] to download samples and prepare dataset of $3,594$ unique samples labeled to 12 state-sponsored APT groups. The dataset is made publicly available [59] for further research.

The study of Gray et al. [42] introduces APTClass, a dataset containing $17,513$ labeled malware samples prepared for research in malware-based attribution. APTClass includes ELF, Windows EXE, and DLL files, prepared by extracting text from PDF reports, YARA rules, and IoC files, performing hash searches for sample collection, and using metadata N-gram and keyword searches for APT label extraction. The dataset can be enriched with OSINT indicators and is available at [63].

TABLE IV: Comparison of the datasets

Dataset	Type	No. of sample	No. of APT group	No. of groups more than 25+ samples	No. of groups more than 50+ samples
APTMalware [31]	Malware	4, 449	12	12	11
APTClass [42][*]	Malware	17, 513	275	85	53
APTNote [19]	Threat Report[a]	687[b]	–	–	–
dAPTaset [26]	Heterogeneous	8, 927	88	14	10
APT Dataset [116]	Malware	1, 200	5	5	5
AttackAttribution [48]	Threat Report	238	12	2	1
Edie et al. [58]	Attack pattern	1, 316	147	0	0
ATT&CK [118]	Attack Pattern				
APT Cyber Criminal Campaign Collection [32]	Threat Report[a]	1, 563[b]	–	–	–

[a] This is the curated list of unlabeled threat reports details along with threat report in pdf format.
[b] This dataset is continuously being updated. So, the number is reported based on current samples in the dataset.
[*] Available upon request. The more details can be found at https://s3lab.isg.rhul.ac.uk/project/aptclass/

Haddadpajouh et al. [116] present a dataset of 1, 200 samples from five APT groups (APT1, APT3, APT28, APT33, and APT37) collected from various public repositories. They leverage the cuckoo sandbox reports to prepare multiple views for each sample, focusing on opcode, bytecode, system call, and header information. This approach captures a multi-dimensional representation of the malware, allowing for an in-depth analysis across several facets. This dataset is made publicly available [69] for the benefit of the community.

2) Threat Report Dataset: Research has also been done using past incident threat reports to perform threat attribution, such as [48], [107]. We explore collection of threat reports available publicly and categorize them in the threat report dataset type. The objective of such dataset is to collect attack information, including attack flow, used tools and techniques, and modus opernadi of threat actors to let model profile threat actors behaviour and aids in threat attribution.

Our exploration find a frequently used threat report collection of APTs named APTNotes [19]. It constitutes an organized repository that aggregates a diverse range of public threat reports, whitepapers, and articles dedicated to APT campaigns range from 2008 to 2023. The repository provides direct links and also maintains archived versions of these resources.

We find a repository APT Cyber Criminal Campaign Collection [32], a vital resource for cybersecurity research. It offers an extensive archive of APT group and cybercriminal campaign information, centralizing documented incidents for a comprehensive understanding of cyber threats and adversary tactics from 2008 to present. This repository aids researchers and cybersecurity professionals by providing open access to detailed analyses of cybercriminal activities to support the development of defense strategies. This repository is up-to-date and continuously being updated by esteemed contributors.

3) Attack Pattern Dataset: Researchers have also leveraged the modus operandi of threat actors in terms of TTPs to profile the threat actors and perform attribution such as [8], [40], [133]. The frequent list of TTPs represents the common patterns a specific threat actor exhibits. We classify such APT group's TTP dataset as the attack pattern dataset.

Edie et al. [58] provide a comprehensive list of extended threat playbooks, i.e., a set of TTPs that an APT group regularly uses. They use association rule mining with the aid of the apriori algorithm to learn the statistical relationships between TTPs. For example, if T1027 and T1059 are observed, then T1105 is hypothesized to exist in the same attack campaign. The hypothesized TTP helps analysts hunt for unobserved TTPs and draw analytical conclusions around attack scenarios. This dataset contains 1, 316 observations and publicly available [75] for community benefits.

MITRE ATT&CK framework [118] also lists a set of TTPs used by threat groups in past attributed attack campaigns. We can map the TTPs associated with a specific threat group through ATT&CK framework [80].

4) Heterogeneous Dataset: This dataset is the collection of numerous data types and their corresponding APT groups, which has been leveraged to propose threat attribution process [66], [70], [83], [114], [119].

Noor et al. [70] developed a dataset derived from publicly accessible CTI documents from leading IT and cybersecurity firms such as Mandiant [67], Unit42 [71], and Crowdstrike [78]. This dataset captures low-level IoCs like malware hashes and IP addresses from CTI reports on various cyber threat actors. Additionally, it incorporates high-level IoCs from the MITRE ATT&CK framework, which details adversaries' TTPs. The creation process involved manually extracting low-level IoCs and synthesizing high-level IoC data to include missed or inaccurately attributed IoCs. This synthesis aimed at a more accurate cyber threat attribution by allowing multiple instances per cyber threat actor. The resulting dataset, comprising low-level and high-level IoCs, seeks to improve cyber threat attribution performance and is available for further study [84].

An Excel sheet focused on APT groups and their operations [103] likely is a critical asset for cybersecurity analysts and researchers. It encompasses a wealth of data, including identifying APT groups, their geographical or organizational origins, details on TTPs employed, associated malware tools, descriptions of known campaigns, and IoCs. This organized compilation facilitates a comprehensive understanding of the threat landscape posed by these sophisticated actors, detailing their methodologies, and the specific vulnerabilities they exploit. By aggregating and categorizing this information, the

document aids in identifying, analyzing, and mitigating threats from various APT groups, enhancing the overall cybersecurity posture against such entities.

MITRE curates the extensive details related to activities performed by threat groups [106]. It endeavors to map APT groups to their TTPs [118], as well as the softwares [111] attackers utilize and the campaigns [115] they conduct, and various naming aliases which acknowledge that different organizations may track the same entities under different names. This approach aids in understanding the overlaps and distinctions among groups, enhancing analyst awareness and encouraging further investigation. Software in MITRE framework encompasses a variety of applications, ranging from custom malware to commonly used operating system utilities, categorized into malware and tools based on their intended use by adversaries. This categorization helps understand the specific techniques that different software instances can employ, mapped to the groups known to use them. Campaigns are defined as structured series of intrusion activities with specific objectives, targets, and duration. MITRE records these activities, providing unique identifiers for unnamed operations or using publicly reported names, which aims to attribute them to known groups and their used software where possible. MITRE facilitates a comprehensive view of threat actor's operations over time and offers insights into the evolution of their tactics and methodologies.

Laurenza and Lazzeretti introduce "dAPTaset" [26], a structured database for APT information. The database structure includes APTs, keywords, threat reports, samples, network information, CVEs (Common Vulnerabilities and Exposures), techniques, and more. The data collection process involves multiple sources such as MITRE ATT&CK [118], APT groups and operations spreadsheets [103], APTNotes [19], and MISP galaxy cluster [13], from which APT names, aliases, victims, operations, and external references are extracted. They parsed the public threat reports to gather IoCs, file hashes, IP addresses, network URLs, emails, CVEs, and keywords related to APTs. The dataset is further enriched by cross-referencing with cybersecurity websites like VirusTotal [16], VirusShare [27], Malshare [34], and Malpedia [47] to validate and supplement information. This dataset uses 821 reports of 88 different threat groups. In addition, it also includes a total of $21,841$ binary hashes, $2,620$ network-related IoCs, $8,927$ unique files, 169 CVEs, and 175 different techniques related to APT activities. The dataset is publicly available [91] for the benefit of the community.

The discussed datasets are crucial for gathering threat actors' characteristics and developing robust attribution methods. Using these datasets, researchers can cross-reference and validate threat intelligence, identify patterns, and attribute cyber incidents to specific threat actors with higher precision. Integrating these diverse data sources is essential for developing robust threat attribution methodologies. A comparison of discussed datasets is present in Table IV.

C. Classification of Attribution methods

We classify the attribution methods in the literature based on dataset and algorithm types. Based on the type of algorithm, we classify the present attribution methods into classification-based, clustering-based, and similarity-based. The classification-based methods are supervised ML methods trained on the labeled datasets, whereas clustering-based methods, i.e., unsupervised methods, are developed on the unlabeled datasets. Additionally, some work is based on matching the similarity between a given sample set with an established threat group database set rather than training any ML model. Based on the type of dataset, we classify available attribution literature methods into four categories: Malware-based methods, Threat report-based methods, Attack pattern-based methods, and Heterogeneous-based methods. Malware-based methods leverage malware samples to perform attribution, Attack pattern-based methods leverage TTP to link threat actors, Threat report-based methods use threat intelligence reports' texts to perform attribution, and Heterogeneous-based methods consider diverse data types to establish the connection between given data and threat actor.

TABLE V: Classification of automated attribution methods

Base	Type	Paper
Algorithm Type	Classification-based	[8], [14], [22], [29], [48], [66], [70], [72], [79], [83], [85], [94], [107], [116], [119]
	Clustering-based	[53], [114], [131]
	Similarity-based	[40], [133]
Dataset Type	Malware-based	[14], [22], [29], [53], [72], [79], [85], [116]
	Threat report-based	[22], [48], [83], [107]
	Attack pattern-based	[8], [40], [114], [131], [133]
	Heterogeneous data-based	[66], [70], [94], [119]

Following the classification based on the dataset type, we comprehensively discuss the literature mentioned in Table V. The details of the literature methods are explained below:

1) Malware-based methods: To leverage malware for threat attribution, various researchers analyze malware's several features and ML models to perform threat attribution [14], [22], [29], [53], [72], [79], [85], [94], [116], [131]. We review current malware-based threat attribution literature and discuss them in detail below.

Xu et al. [14] present a combined feature set of TF-IDF and bi-gram methods prepared by API call sequences of APT samples. The combined features are huge in number, so they employ the Adaboost feature selection method to rank the features based on their importance and contribution to classification. They select the top-400 feature set as input and deploy several multiclass ML classifiers, i.e., DT, KNN, LightGBM, and XGBoost, to classify malware to their corresponding APT

groups. Their result finds LightGBM as the best performer compared to all implemented classifiers.

Kida et al. [29] present a novel way to leverage malware's fuzzy hashes to attribute APT groups. They calculated five types of fuzzy hashes: TLSH, SSDEEP, SDHASH, IMPFUZZY, and LZJD for all malware samples and used computed hash as natural language input to the classifier. After hash calculation, the hashes are encoded using the NATO encoding method, and they use the padding method to keep the encoded length the same. Further, they use character level and n-gram vector embeddings to convert encoded hashes into feature vectors. The embedding vectors are forwarded to four different ML classifiers: RF, SVM, KNN, and NB. Their result demonstrates that RF with IMPFUZZY hash performs best of all implemented methods.

Rosenberg et al. [79] propose a method named DeepAPT to employ raw dynamic malware features and attribute APT groups linked with the given malware sample. They execute samples in a sandbox environment and collect behavior reports for each malware sample. DeepAPT considers words present in behavior reports as features and evaluates the frequency of each unique word in all behavior reports. The authors selected the top $50,000$ frequent words as features and evaluated one-hot encoding to prepare the dataset. Further, they employ a 10-layer DNN to learn patterns from raw dynamic features for APT group classification.

Wang et al. [22] introduce a novel approach for APT attribution by combining code and string features using paragraph vectors and bag-of-words vectors, respectively. They employ a RF classifier and DNN for classification by leveraging both types of features. Further, they add model interpretation techniques using Local Interpretable Model-agnostic Explanations (LIME), which offer insights into the classification process, enhancing trust and facilitating analysis for cybersecurity personnel.

Li et al. [85] propose an ML-based classification method for attributing APT malware to specific organizations in the IoT. Their method analyzes samples dynamically and extracts features from the dynamic analysis report. They extract the first $10k$ words and weigh each word using the TF-IDF method to transform features into vector form. Their method also includes filtration of stop-words before feature transformation. Out of all behavioral features, they employ the chi-square method to select features and use SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset. Further, they leverage the RF classifier to perform multiclass classification between APT groups for a given malware sample.

Wang et al. [53] introduce a novel method for automating the attribution of APTs by analyzing binary code. Their method identifies and classifies APT malicious codes by utilizing local features within binary functions and employing time series mining techniques. They extract 18 different features of basic blocks from the CFG of the executable. They also filter the non-relevant functions to focus on those containing API calls indicative of malicious behavior. Next, they generate paths of interest based on API calls and train the PCA (Principal Component Analysis) to reduce the dimension from 18 to 1. They use HDBSCAN (Hierarchical Density-Based

Spatial Clustering of Applications with Noise) to choose paths based on cluster classes and generate time series shapelets to represent key subsequences of paths. Next, they employ an RF classifier to attribute APT organizations to malware samples based on the distances between paths and path shapelets.

Haddadpajouh et al. [116] introduce a multi-view fuzzy consensus clustering model to attribute cyber threat payloads, specifically malware, to their respective actors. Their study follows a multi-view approach (multiple pieces of evidence) rather than single-view (single piece of evidence) analyses. The multi-view is generated using four categories of features: Header, System call, Byte code, and Opcode. The authors calculate multi-view for each feature category using binary vector (1 if present 0 if absent), count vector (number of it appears), frequency of occurrence, TF-IDF vector, and eigenvector of CFGs. Further, they utilize fuzzy pattern trees and multi-modal fuzzy classifiers for inference. Their research demonstrates that a multi-view approach is significantly better than single-view analyses for cyber threat attribution.

Kim et al. [131] present a method for automatically attributing mobile APT groups by leveraging a vectorized ATT&CK matrix and IoC pairing. It involves two primary steps: first, collecting and mathematically modeling TTPs from the MITRE ATT&CK framework, followed by vectorization and storage for similarity measurement; second, enhancing attribution accuracy by comparing IoC pairs to counteract potential false flags by attackers. They use Joe Sandbox to extract ATT&CK TTPs from malware. The experiments involve comparing the cosine similarity of the ATT&CK matrix TTPs and the IoC pairings among the malware samples.

Lee and Cho [72] propose a model for malware authorship attribution that leverages runtime modules to attribute linked APT groups. They perform frequency analysis to execute the samples to collect runtime-module names like Kernel32.dll and Advapi32.dll. They leverage one-hot encoding to convert features into numerical vectors and feed the vector to a set of ML classifiers for classification. They find the XGBoost algorithm as the best-performing model for attribution.

Mei et al. [94] present a hybrid approach combining Particle Swarm Optimization (PSO) and Multiclass SVM (MSVM) for APT organizations by analyzing traces of APT attack tools in a sandbox environment. The approach optimizes MSVM parameters for accurate APT organization identification. They consider hybrid information as feature set, including file_type, malware_type, label, size, target_object, etc. They use the TF-IDF method to prepare feature vectors and PSO method to initialize the parameters of MSVM. Further, they train the classification model to identify APT organization.

2) Threat Report-based methods: Some researchers leverage texts of past attack incidents threat reports to distinguish between campaign details of APT groups [48], [107]. We reviewed these methods and discussed them in detail below.

Perry et al. [48] present NO-DOUBT, a novel ML-based method for attack attribution leveraging textual analysis of threat intelligence reports. They propose a new text representation algorithm, SMOBI (Smoothed Binary vector), designed

to minimize feature engineering while capturing the contextual nuances of words through vector space representation derived from a labeled dataset and a vast corpus of security literature. They use report text as a feature, which is better for identifying similar reports than pinpointing threat actor behavior. SMOBI focuses on word frequency and similarity, ignoring semantics and context. Their approach also allows for identifying known and novel threat actors by classifying text into specific actor categories or signaling new, previously unseen actors.

Naveen et al. [107] introduce a deep learning architecture leveraging DNNs and domain-specific word embedding for attributing threat actors based on threat intelligence reports. The novel vector representation method, SIMVER, utilizes semantic relationships within the text to improve attribution performance significantly. They tested their method against traditional methods using a dataset of threat reports [107]. Their method SIMVER demonstrates better performance in accurately identifying threat actors.

3) Attack Pattern-based methods: Various researchers have used threat actor's attack patterns in terms of TTPs to profile threat groups and perform attribution [8], [40], [133]. We discuss these methods in detail below.

Noor et al. [8] propose an automated framework for cyber threat attribution by profiling threat actors based on their high-level attack patterns extracted from CTI reports. They develop a semantic search system based on the statistical distributional semantic relevance technique and map them according to ATT&CK framework TTPs. Further, they prepare a feature correlation matrix for each threat report and employ ML classifiers to classify threat groups. They also observe the impact of features using the Information Gain (IG) feature selection method.

Edie et al. [40] present a novel approach to cyber attack attribution using association rule mining and a weighted Jaccard similarity index. Their approach involves learning associations between TTPs using MITRE ATT&CK framework. These associations are represented as activity groups (AGs) using the apriori algorithm, aiding in identifying statistically related TTPs. The authors generate the standard and extended threat playbooks by incorporating confidence levels for known and hypothesized TTPs. To identify potential adversaries, the algorithm is seeded with observed TTPs. Further, they use a weighted Jaccard similarity index to compare observed TTPs with threat playbooks and attribute attacks to likely adversaries.

Shin et al. [133] introduce Automated Reclassification for Threat Actors (ART) to automatically compare the TTPs from different APT groups by leveraging the MITRE ATT&CK framework. Their method involves crawling cyber threat reports and retrieving the ATT&CK matrix of APT groups, preparing TTP-vectors, and calculating the cosine similarity to re-examine and classify the APT groups. They prepare 12 different one-hot encoding vectors for each tactic in the ATT&CK framework (ignores Pre-ATT&CK tactics). Further, they calculate cosine similarity tactic-vector-wise between all groups and employ a weighting mechanism across the similarity score of all 12 tactic vectors.

4) Heterogeneous data-based methods: Along with the single type of data, i.e., malware, TTP, or threat report, several researchers have leveraged multiple types of data merged to perform attribution [66], [70], [83], [114], [119]. They merge all relevant data about a threat actor and then use it as a dataset to develop a method for attribution. We discuss all of them in detail below.

Noor et al. [70] leverage two types of high-level IoCs, i.e., TTPs and Software used by threat actors, to perform attribution. They leverage the list of TTPs present in ATT&CK framework [118] and software provided by MITRE [111] to prepare the dataset. They extract these high-level IoCs from threat reports and prepare feature vectors using the one-hot encoding method. Further, they employ several ML and DL classifiers to perform attribution and select the best-performing classifier as the final attribution model. They also compare attribution performance based on high-level and low-level IoCs and find that high-level IoCs are more trustworthy than low-level IoCs for threat attribution.

Sachidananda et al. [119] present methods named APTer, an innovative framework designed to tackle the complexity of APTs by correlating, predicting, and attributing stages of APT attacks, as well as mapping CVEs to the MITRE ATT&CK framework. Their multifaceted approach includes eliminating redundant alerts, clustering related alerts to identify attack scenarios, and using ML models to predict future attack stages and attribute attacks to known APT groups based on their TTPs. The attribution module of their proposed method implements an ML classifier on the set of TTPs seen during the attack and performs attribution classification to find linked threat groups. They extract TTPs from heterogeneous threat alerts, i.e., IDS/IPS, SIEM, and Firewall. Their data modeling stage completes the partial TTP view by predicting the next attack stage and completing the missed TTP based on the cyber kill-chain phase across the MITRE tactics.

Xiao et al. [83] introduce an APT actor attribution method named APT-MMF, focusing on multimodal and multilevel feature fusion from CTI reports. This method employs multilevel heterogeneous graph attention networks that integrate IoC type-level, meta path-based neighbor node-level, and meta path semantic-level attention to learn deeply hidden features of APT report. Further, they extract and combine three features: attribute types, natural language text, and topological relationships. For natural language text features, they employ BERT encoding [128], and for topological relationship features, they utilize Node2vec [135] encoding, and for attribute type features, they use ID encoding and Ordinal encoding. Further, they implement GNN to perform attribution.

Xu et al. [66] present a method named HGHAN, designed for identifying hacker groups through a heterogeneous graph attention network. The HGHAN framework first extracts relevant node sets and attributes from the Web Attack Heterogeneous Information Network (WAHIN), followed by applying LSTM for attribute feature extraction and Metapath2Vec for structural feature extraction. The fused features are subsequently processed through HAN (Heterogeneous Graph Attention Network) for node embedding, with classification being carried out using BiLSTM to leverage its proficiency

in capturing bidirectional dependencies within the feature vectors. They conducted experimental validation using web attack data from Zone-H.org to demonstrate the performance of HGHAN over other node embedding algorithms.

Wu et al. [114] introduce GroupTracer, a framework aimed at identifying and predicting Internet of Things (IoT) attacks by extracting TTP profiles that describe attacker behaviors and identifying potential attacker groups behind complex attacks. GroupTracer captures attack data via IoT honeypots and utilizes the ATT&CK framework to map and extract TTP profiles automatically. It considers total 18 different features derived from four feature category: TTP profiles, Time, IP, and URL features-to mine potential attack groups. They leverage the hierarchical clustering algorithm's ability to capture the hierarchical structure between clusters.

A comparison of used features, their transformation, and the algorithm employed by the discussed methods are shown in Table VI. For malware-based methods, authors in literature extract the following types of features from malware samples: words from cuckoo sandbox reports, OpCode, ByteCode, API call, Header data, IR (Intermediate Representation) codes, strings, Fuzzy hashes (TLSH, SSDEEP, SDHASH, IMPFUZZY, and LZJD), CFG, and runtime modules. Threat report-based attribution methods extract words from threat intelligence reports as features. Attack pattern-based methods leverage TTPs as a feature for attribution. Meanwhile, heterogeneous data-based attribution methods leverage diverse sets of features like TTPs, software used, and heterogeneous attributed graph schemas. The distribution of feature types used in the literature is shown in Fig. 7a.

Out of all features mentioned in the literature TTPs and threat report texts has been used widely to present attribution method. For feature transformation, the literature leverages the following different methods: One-Hot encoding, SMOBI word embedding, TF-IDF, Count, Statistical Characteristics, SIMVER word embedding, n-gram method, NATO encoding, Basic block characteristics, frequency analysis, Node embedding, BERT embedding, and Node2vec embedding. The distribution of type of transformation methods has been used in the literature is shown in Fig. 7b. Out of all used transformation methods One-Hot encoding and TF-IDF has been used widely in the literature.

Literature methods employ the following algorithms for classification or clustering: DNN, XGBoost, DT, Hierarchical clustering, KNN, LightGBM, RF, k-means clustering, Cosine similarity, SVM, NB, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering, BiLSTM, Weighted Jaccard Index, DNN, Sequential classifier, and GNN. The distribution of type of algorithms used in the literature is shown in Fig. 7c. Out of all employed classification model DNN, DT, and RF has been used widely in the literature to train attribution model.

D. Discussion

The discussed papers demonstrate a range of strengths and weaknesses in their approaches to threat attribution. In this section, we provide a detailed analysis of these aspects, thoroughly examining the methodologies and outcomes of each paper.

We review the literature methods comprehensively and find five major data related problems among the available methods, which we discussed below.

(P1) As malware can be tweaked, distributed, or sold, i.e., given the rise of MaaS (Malware-as-a-Service) [33], it may lead to notable challenges in uncovering the actual origin of a cyber attack. Malwares' behaviour-based threat attribution may provide an inconsistent view of threat actors when the same threat actor employs numerous malware in various campaigns. **(P2)** Malware-based threat attribution can be misleading in the case of collaborative and multi-layered attacks, as these attacks do not rely on a single instance of malware. **(P3)** For the threat-report-based attribution methods, authors employ threat report words as features for classification in the literature. Comparing the words of two threat reports shows the similarity between threat reports rather than the similarity between the threat actor's behavior or modus operandi. Comparing words for document similarity has been studied widely in the natural language processing domain [28], [38]. **(P4)** Given P3, such methods need to reverse engineer and identify words contributing to the attribution or profiling of a specific threat actor. **(P5)** Attack pattern-based methods leverage TTPs in the literature. TTP of a threat group can evolve over time [52], and TTP-based threat attribution can miss-attribute in such cases. Therefore, TTP-based attribution methods demand continuous updates according to the threat group's latest modus operandi. Also, threat actors may employ false flags by employing other groups of TTPs to disrupt the attribution method or hide their traces [21], [24].

Further, we provide critical comments on the current literature methods and their reported performance in the respective studies in Table VII. The discussed critical comments appeal to evaluate literature methods and future guidance for new research dimensions in threat attribution. While reviewing these literatures, we also find that the majority of the literature suffers from the following limitations:

(a) **Comparison with State-of-the-art methods:** Majority of the literature lacks the comparison with the state-of-the-art attribution method. The reported performance of literature methods are good, but a comparison with state-of-the-art methods in the same environment and common dataset can establish a comprehensive comparison and prove the efficiency of the proposed methods.

(b) **Limited Threat Group Implementation:** The number of threat groups researchers use to implement their methods is limited, with very few APT groups. For example, the method of [116] is tested on only five APT group samples, and the method of [133] is tested on only three APT group samples. Such limited environment testing raises concerns regarding method effectiveness for diverse and complex real-world data.

(c) **Close-world Solutions:** APT groups evolve over time, and several threat groups emerge nowadays. It raises concerns regarding models trained on close-world solutions. Close-world solutions attribute threat groups to only what

TABLE VI: Comparison of automated attribution methods, their features, and the algorithms employed in each method

Paper	Year	Artifact Type	Method Type	Chosen Feature	Feature Transformation	Algorithm
[79]	2017	Behavioral Artifacts	Malware-based	Words of cuckoo sandbox reports	One-Hot encoding	DNN
[48]	2019	OSINT Artifacts	Threat Report-based	Words of threat intelligence reports	SMOBI word embeddings	XGBoost
[8]	2019	Behavioral Artifacts	Attack Pattern-based	TTPs	One-Hot encoding	DNN
[116]	2020	Behavioral Artifacts	Malware-based	OpCode, ByteCode, System call, and Header	One-Hot, Count, TF-IDF, and Eigen vector of CFG	DT
[114]	2020	Behavioral and Evidentiary Artifacts	Attack Pattern-based	Time, IP, URL, and TTPs mapped from IoT honeypot logs	Statistical characteristics for each feature types	Hierarchical clustering
[107]	2020	OSINT Artifacts	Threat Report-based	Words of threat intelligence reports	SIMVER word embeddings	DNN
[14]	2021	Behavioral Artifacts	Malware-based	API call sequence	n-gram with Adaboost feature selection	DT, KNN, LightGBM, and XGBoost
[22]	2021	OSINT and Behavioral Artifacts	Threat Report and Malware-based	IR code and string features	Paragraph and bag-of-words vectors	RF and DNN
[85]	2021	Behavioral Artifacts	Malware-based	Words of cuckoo sandbox reports	TF-IDF with Chi-square for dimensionality reduction	SMOTE and RF
[131]	2021	Behavioral Artifacts	Attack Pattern-based	TTPs	Tactics-wise One-Hot encoding	k-means clustering
[133]	2021	Behavioral Artifacts	Attack Pattern-based	TTPs	Tactics-wise One-Hot encoding	Cosine similarity with weighted tactic vectors
[29]	2022	Behavioral Artifacts	Malware-based	TLSH, SSDEEP, SDHASH, IMPFUZZY, and LZJD	NATO encoding	RF, SVM, KNN, and NB
[53]	2022	Behavioral Artifacts	Malware-based	Control Flow Graph (CFG)	Basic blocks characteristics with PCA for dimensionality reduction	HDBSCAN clustering
[72]	2022	Behavioral Artifacts	Malware-based	Runtime Modules like Kernel32.dll	Frequency analysis and One-Hot encoding	KNN, SVM, DT, NB, Adaptive and Gradient Boosting
[94]	2022	Behavioral and Evidentiary Artifacts	Heterogeneous Data-based	Log data	TF-IDF	Multi-class SVM with Particle Swarm Optimization
[66]	2022	Evidentiary and OSINT Artifacts	Heterogeneous Data-based	Web attack heterogeneous graphs	Node embedding	BiLSTM
[40]	2023	Behavioral Artifacts	Attack Pattern-based	MITRE TTP and it's association	TTP List	Wighted Jaccard Index
[70]	2023	Behavioral Artifacts	Heterogeneous Data-based	TTPs and Softwares used	One-Hot encoding	Neural Network
[119]	2023	Behavioral and Evidentiary Artifacts	Heterogeneous Data-based	TTPs from IDS/IPS and SIEM/EDR Alerts	TTP list	Sequential classifier
[83]	2024	OSINT Artifacts	Threat report-based	Node features of heterogeneous attributed graph schema	ID encoding, Ordinal encoding, BERT encoding, and Node2vec encoding	GNN

they have seen during training. The emergence of threat groups leads to misattribution when the given sample belongs to newly emerged groups whose sample has not been seen by the model during training. As a result, the trained model is attributed to a wrong threat group. Our review also finds that only [48] introduced the open-world

(a) Feature types   (b) Feature transformation methods   (c) Algorithm types
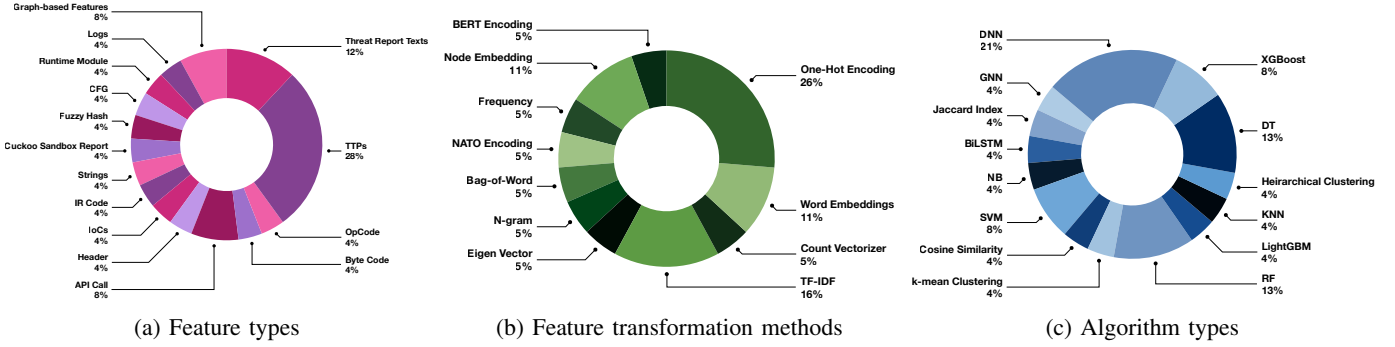
Fig. 7: Distribution of feature types (a), transformation methods (b), and algorithm types (c)

solution, i.e., capability of their method for identifying unknown or newly emerged threat actors.

(d) **Lacks Contextual Information:** The current attribution literature is majorly based on the existence of a feature rather than leveraging its contextual information. It is crucial to see how an artifact is being used during the attack to connect the dots with historical threat groups' behavior [35]. For example, TTP X is used to gain the initial access, and subsequently TTP Y is used for privilege escalation. This sequential information may hint at more aligned behavior rather than just matching the appearance of the TTPs individually. Most of the literature leverages the One-Hot encoding and TF-IDF feature transformation method to convert selected features into vectors, which does not consider contextual information. Some of the work has leveraged contextual information of the artifacts to perform attribution [14], [66], [83], [119]. Using context-aware feature transformation methods or consideration of contextual information can assist researchers in developing attribution methods that make informed decisions.

(e) **Granularity Level:** Granularity in attribution can significantly influence the effectiveness and reliability of the attribution process. It refers to the level of detail and precision with which cyber attacks are attributed to their perpetrators. According to 4C model [7], the granularity of attribution is divided into four levels: Technical, Motivation, Country, and Organization/Person. Technical attribution involves identifying and analyzing technical artifacts like malware signatures and IP addresses. Motivation attribution focuses on understanding the underlying reasons for the attack, such as financial gain, espionage or disruption. Country attribution aims to determine the geographical origin of the attackers by examining factors like language settings and geopolitical interests. The most detailed level, Organization/Person attribution, involves identifying specific organizations or individuals responsible for the attack, often requiring extensive investigation and evidence gathering. Each level provides a deeper understanding of the attack, contributing to a comprehensive attribution process. Literature like [79] focuses on attributing only nation rather than the specific threat group. The attribution demands uncovering

as many specific details as possible about the threat actor. Therefore, fine-grained attribution allows for more tailored and effective defensive measures, enhanced intelligence sharing, and informed strategic decision-making.

(f) **Performance Metrics:** The performance measured in the current literature is dispersed. Some of the current methods [22], [40], [79], [116] compare the efficiency of literature models based on accuracy measures, Whereas some of the literature focus on other metrics like precision, recall, or f1-score, such as [53], [114]. The goal for the specific case of attribution is to get better link between given artifact and associated threat group. Along with this, incorrect attribution can result in misdirected cybersecurity efforts, wasted resources, and possible diplomatic conflicts, weaken the credibility of the attribution process, escalate tensions, and may lead to incorrect strategic decisions [127]. Therefore, the false positive needs to be minimized, i.e., precision is more significant than other performance metrics in threat attribution.

VI. CHALLENGES

Attribution itself is a challenging and complex task and requires diverse consideration during the process. Additionally, APT attacks' sophisticated, multifaceted, multi-staged, and dynamic nature makes it more difficult. The challenges faced by APT attributions are discussed in this section.

A. Lack of Comprehensive and Structured Data

A significant challenge in APT attack attribution is the lack of comprehensive and structured datasets. The available threat intelligence is often incomplete or fragmented, making it difficult to form a complete picture of the threat landscape. This fragmented data can lead to partial or incorrect conclusions, as investigators might miss critical information necessary for accurate attribution. Our presented artifact taxonomy serves as a foundational step toward creating structured data for attribution. Researchers can utilize this taxonomy to identify and collect the relevant data types necessary for attribution. Furthermore, the quality and reliability of threat intelligence data vary widely, with some sources providing high-fidelity data while others contribute less reliable information. This variability further complicates the accuracy of attribution efforts, as inconsistent and unreliable data can result in incorrect

attributions [92]. Without robust and structured datasets, the attribution process remains severely hindered, potentially leading to flawed assessments and ineffective responses to cyber threats.

TABLE VII: Comparison of attribution methods and their performance

Paper	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Context-aware	Source Code	Open World	Granular	Comments
[79]	94.60	—	—	—	✗	○	✗	✗	• Attributes at the country level rather than the threat group level • Word semantics of Cuckoo report reflect malware functionality rather than threat actor behavior • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[48]	58.40	55.00	52.40	—	✗	◑	✓	✓	• Their feature vector method, SMOBI, focuses on word frequency and similarity, ignoring semantics and context • As the method is leveraging words present in the threat report to attribute threat groups, it faces problems **P3** and **P4** as discussed in the Section V-D
[8]	94.00	90.00	89.00	89.00	✗	○	✗	✓	• Model is trained on the highly imbalanced and limited dataset • Their TTP extraction method is based on semantic similarity rather than the contextual meaning of sentences • The feature transformation method considers only the existence of TTPs and ignores contextual information such as when and how TTP has been used
[116]	95.00	—	—	—	✗	◑	✗	✓	• Dataset used is limited to five APT groups; lacks diversity. A larger and more diverse dataset would strengthen the findings • Claims significant improvements but lacks comparisons with the latest state-of-the-art methods in malware-based attribution • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[114]	—*	—*	—*	—*	✗	○	✗	✓	• Dataset used is limited to specific honeypots (UPnP-SOAP and Netis router backdoor) • Proposed method compare with two traditional clustering method but lack the comparison with state-of-the-art attribution methods
[107]	86.50	95.40	83.30	87.90	✗	○	✗	✓	• The feature transformation method, SIMVER, focuses on word semantics but ignores the contexts of sentences • As the method is leveraging words present in the threat report to attribute threat groups, it faces problems **P3** and **P4** as discussed in the Section V-D
[14]	93.53	89.65	84.72	86.62	✓	○	✗	✓	• Computationally intensive with large feature sets. Despite selecting the top 400 features, scalability remains a concern with larger dataset • Detail comparison with various ML classifiers but lacks comparison with state-of-the-art malware-based attribution methods • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[22]	97.87	—	—	—	✗	○	✗	✓	• Rely on static features for the decision which is challenging in the case of obfuscated samples. The obfuscation is common in APT types of attacks • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D

Continue on the next page

● Source code is available, ◑ Only data is available, ○ No source code is available
— Not Available
—* Calinski–Harabasz score: 3416.9311, Silhouette Coefficient: 0.5389, Davies–Bouldin: 0.7367

TABLE VII: Attribution Methods and their Performance Comparison

Paper	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Context-aware	Source Code	Open World	Granular	Comments
[85]	–	85.00	79.10	79.40	✗	○	✗	✓	• No evaluation to demonstrate the data generated by SMOTE aligns with base data and not contributing towards biases • No comparison with state-of-the-art malware-based threat attribution methods • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[131]	–	91.48	95.14	93.27	✗	○	✗	✓	• IoC pairing method can mislead in false flags as chosen IoCs like C&C Platform, their IP and malware types, can be modified or changed in APT attacks • No comparison with state-of-the-art malware-based threat attribution methods • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[29]	89.00	–	–	88.00	✗	○	✗	✓	• The authors reported their method struggles to surpass the state-of-the-art, aiming to understand how fuzzy hashing aids in identifying nation-state actors • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[53]	–⊛	–⊛	–⊛	–⊛	✗	○	✗	✓	• Relies on static features, which can be less effective against advanced sophisticated malware which employs various obfuscation methods • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[94]	93.96	93.97	–	93.94	✗	○	✗	✓	• No comparison with state-of-the-art malware-based threat attribution methods • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D
[40]	95.70	–	–	–	✗	●	✗	✓	• Reporting only accuracy is insufficient; additional metrics like precision, recall, and F1-score are needed for a comprehensive evaluation. • Lacks a thorough comparison with existing state-of-the-art methods for APT attribution • As the method is leveraging TTPs as feature for attribution, it faces problem **P5** as discussed in the Section V-D
[70]	94.88	93.95	94.88	94.00	✗	◑	✗	✓	• Lacks a thorough comparison with existing state-of-the-art methods for cyber threat attribution • While preparing dataset, the process of synthesis of the high-level IoC dataset is not explained
[119]	97.30	96.00	97.00	97.00	✓	○	✗	✓	• While preparing dataset, mapping SIEM/EDR alerts to MITRE ATT&CK TTP is done by a security tool but there is no mention of the tool name • The performance of the method is depends on the mapping efficiency of chosen security tool using which data is prepared • As the method is leveraging TTPs as feature for attribution, it faces problem **P5** as discussed in the Section V-D
[83]	–	–	–	70.51	✓	○	✗	✓	• Lacks a thorough comparison with existing state-of-the-art methods for cyber threat attribution • The accuracy and efficiency of analytical calculations depend on the lengths and complexity levels of metapaths. There is no evaluation presented regarding computational complexity

● Source code is available, ◑ Only data is available, ○ No source code is available
− Not Available
−⊛ AUC (Area Under the Curve): 93.6%

TABLE VII: Attribution Methods and their Performance Comparison

Paper	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Context-aware	Source Code	Open World	Granular	Comments
[66]	98.00	97.00	98.00	97.00	✓	○	✗	✗	• No comparison with state-of-the-art methods • Dataset used is limited to web-based application, i.e., Web-Hacking dataset
[72]	87.10	89.00	87.00	87.0	✗	○	✗	✓	• Dataset used is limited to only three APT group • As the method is leveraging malware sample to attribute threat groups, it faces problems **P1** and **P2** as discussed in the Section V-D

● Source code is available, ◖ Only data is available, ○ No source code is available
− Not Available

B. False Flags

APT groups may deliberately leave false clues, known as false flags, to mislead investigators about their true identity [10], [24]. This deceptive tactic involves using tools, techniques, and methods typically associated with other groups or countries [21]. By adopting these recognizable signatures, they aim to create confusion and misdirection, making it appear that another entity is responsible for the attack. This strategy complicates the attribution process and increases the likelihood that investigators will draw incorrect conclusions about the origin and perpetrators of the cyber assault. False flags are the sophisticated method to obscure the true actors behind an attack, highlighting the complexity and challenges in accurate cyber threat attribution.

C. Lack of robustness against coordinated attack

Attribution efforts face significant challenges due to the attack-latent relationships between APT groups. Some APT groups collaborate against common victims to conduct joint operations and strive toward mutual strategic goals [92]. This collaboration complicates the attribution process, blurring the lines between different groups' activities. When APT groups work together, it becomes challenging to determine which group is responsible for a specific attack. The overlapping use of shared resources and coordinated strategies can mislead investigators, making it difficult to pinpoint the attack's origin and attribute it to the correct entity. This complexity underscores the need for advanced and nuanced attribution methods capable of disentangling these collaborative efforts.

D. Shared Tools

Publicly available tools are commonly used by many APT groups [42], making it challenging to differentiate between actors solely based on tool usage. For example, both APT29 and Kimsuky utilize PsExec, though they are reportedly sponsored by different nation-states [42]. These shared tools and techniques among various threat actors further blur the lines, making it difficult to attribute an attack to a specific group solely based on technical indicators. When multiple groups use the same tools, it becomes challenging to identify unique signatures or patterns that could definitively link an attack to a particular group. This widespread use of shared resources complicates the attribution process, as investigators cannot rely on tool usage as a clear differentiator. Thus, the shared nature of these tools necessitates more sophisticated methods to attribute cyber attacks to the correct perpetrators accurately.

E. Anonymization

APT groups often employ sophisticated anonymization techniques, such as Tor, Virtual Private Networks (VPNs), and proxy servers, to mask their actual IP addresses and geographical locations [54]. Using these methods, they can effectively conceal their origins, making it exceedingly difficult for investigators to pinpoint the trustworthy source of an attack. These deliberate misdirections complicate the attribution process, as they create layers of obfuscation that need to be penetrated to identify the attackers accurately. The presence of such anonymization necessitates advanced forensic techniques and extensive analysis to discern the real origin of the threat. Automated systems often struggle to achieve this level of performance, and highlights the importance of advanced analytical methods in attributing sophisticated cyber attacks.

F. Interdisciplinary Nature

The interdisciplinary nature of cyber threat attribution adds another layer of complexity. Accurate attribution requires integrating technical expertise in cybersecurity, malware analysis, and network forensics with non-technical insights from geopolitics, sociology, and intelligence analysis [68]. This interdisciplinary approach is necessary to understand an attack's broader context and motivations. However, achieving such integration is challenging and requires collaboration between stakeholders, including private companies, governments, and international organizations.

G. Legal and Ethical Implications

The legal and ethical considerations of cyber threat attribution are essential and must be diligently addressed [51]. Collecting and analyzing data for attribution can infringe on privacy rights and raise significant ethical concerns [117]. Incorrect or premature attribution can lead to severe legal and diplomatic repercussions, potentially escalating conflicts and causing harm to innocent parties. These considerations necessitate a cautious and methodical approach along with automated systems. Balancing the need for accurate and timely attribution with the ethical and legal ramifications adds to the overall difficulty of the task.
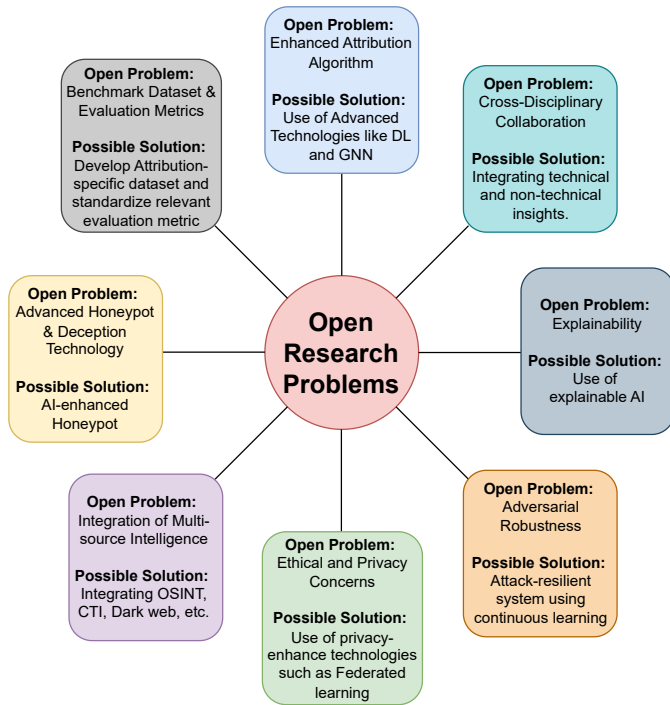
Fig. 8: Open Research Problems and Possible Future Research Directions

VII. OPEN RESEARCH PROBLEMS

The difficulty of automated cyber threat attribution presents numerous open research opportunities and future research directions. As cyber threats evolve in sophistication and complexity, there is a growing need for innovative approaches and technologies to improve the accuracy and reliability of threat attribution. Based on our survey, we present some critical open research problems where future research can significantly advance APT attribution as shown in Fig. 8 and discussed in this section.

A. How does one enhance the current automated APT attribution methodologies?

Automated cyber threat attribution presents numerous research opportunities that promise to advance the field of cybersecurity significantly. Enhancing attribution algorithms through sophisticated ML and AI techniques remains a top priority. By incorporating advanced models such as GNNs, researchers can better recognize patterns within complex cyber attack data, leading to more accurate and efficient attribution. These models excel at identifying intricate relationships and connections within the data that traditional methods might miss. Additionally, context-aware attribution methods are being developed to improve accuracy in considering the broader context of an attack, such as geopolitical factors and historical attack patterns. By integrating these advanced techniques, automated attribution systems can become more robust and reliable, offering a powerful tool for cybersecurity professionals to identify and respond to threats.

B. How can insights from diverse fields be effectively integrated into automated attribution systems to enhance their robustness and depth?

Cyber threat attribution significantly benefits from cross-disciplinary collaboration, combining insights from various fields such as cybersecurity, geopolitics, sociology, and intelligence analysis. This integrated approach ensures that technical data is enriched with a contextual understanding, providing a more comprehensive view of the threats. Researchers can explore automated frameworks and methodologies to foster such collaboration, creating systems where experts from different disciplines can contribute their unique perspectives. This interdisciplinary approach can lead to more robust and nuanced attribution conclusions, allowing for a deeper understanding of cyber attacks' motives, methods, and implications. By leveraging the strengths of diverse fields, the attribution process can become more accurate and insightful, ultimately enhancing overall cybersecurity efforts.

C. How can we create explainable and transparent attribution algorithms that provide clear and understandable reasoning for their decisions to gain trust and facilitate adoption?

The transparency and trustworthiness of AI models used in threat attribution are essential focus areas. Developing explainable AI models is crucial, as these models need to provide clear and understandable reasoning for their decisions. This transparency is vital for gaining user trust, as stakeholders must understand how and why certain conclusions are reached. Explainable AI can help users feel more confident in the technology and ensure its adoption in critical applications. By making AI models more interpretable, developers can bridge the gap between complex algorithms and human understanding to facilitate better decision-making and foster greater acceptance and reliance on AI-driven threat attribution systems.

D. How can we design attribution algorithms robust to adversarial attacks used by sophisticated threat actors to mislead investigators?

As cyber threats evolve, attribution models must enhance their robustness against adversarial attacks. It involves designing models that are resilient to attempts at misleading attribution efforts. Attackers often use sophisticated methods to deceive these models, so creating attack-resilient systems is essential. Additionally, implementing continuous learning mechanisms allows these models to adapt to the ever-changing threat landscape. This adaptive capability ensures that automated attribution systems remain effective over time, maintaining their ability to accurately identify and attribute cyber threats despite the evolving tactics of adversaries. Robust and adaptable models are crucial to sustain the reliability and performance of cyber threat attribution in a dynamic environment.

E. How to develop privacy-preserving techniques to protect sensitive data in attribution processes?

Ethical and privacy considerations are critical in the development of future attribution technologies. Researchers must

create privacy-preserving techniques to ensure that attribution processes do not infringe on user privacy. It involves implementing methods that protect sensitive data while allowing accurate attribution. Additionally, establishing ethical guidelines is essential to govern the responsible use of these technologies. These guidelines will help ensure that attribution efforts are conducted in a manner that respects individual rights and adheres to legal and moral standards. By addressing ethical and privacy concerns, developers can foster trust and acceptance of attribution technologies, ensuring their responsible and effective application in cybersecurity.

F. How can diverse data sources be effectively integrated and harmonized to improve the APT attribution?

Integrating multiple intelligence sources, such as OSINT and traditional CTI, can significantly enhance the understanding of threat actors and their activities. OSINT, which includes publicly available information from various online platforms, can complement traditional intelligence methods by providing additional context and insights. Additionally, mining dark web data for signs of attack planning and correlating it with real-world incidents offers a rich, untapped potential. This approach can reveal deeper insights into cybercriminals' motives and methods, allowing for a more comprehensive and nuanced view of the threat landscape. By leveraging OSINT and dark web data, security professionals can develop a more detailed and accurate picture of cyber threats, improving their ability to anticipate and respond to attacks.

G. What methodologies can be developed to improve the ability of honeypots to detect and capture malicious activities and accurately identify the perpetrators behind these attacks?

Advanced honeypots and deception techniques are emerging as promising research areas in cybersecurity. AI-enhanced honeypots, which can analyze captured attack data and automatically attribute it to specific threat actors, offer significant benefits by providing valuable intelligence with minimal human intervention. These sophisticated honeypots detect and capture malicious activity and help identify the perpetrators behind the attacks. Additionally, developing new deception strategies can further improve the accuracy of attribution data. Security researchers can gather detailed insights into threat actors' behavior, techniques, and tactics by misleading attackers and studying their responses. These advanced methods enhance our understanding of cyber threats and improve the effectiveness of attribution efforts.

H. How do we benchmark datasets and evaluation metrics to rigorously assess and compare the performance of different automated attribution algorithms?

Benchmark datasets and evaluation metrics are essential for advancing the field of cyber threat attribution. Creating standardized benchmark datasets enables researchers to assess and compare different attribution methods rigorously. These datasets provide a consistent basis for testing, ensuring that new techniques are evaluated against a common standard.

Evaluation metrics, particularly those focused on precision, are crucial for measuring the accuracy and reliability of attribution methods. High precision ensures the methods correctly identify actual threat actors while minimizing false positives. Researchers can drive continuous improvement in attribution technologies by emphasizing precision in evaluation metrics, leading to more accurate and trustworthy results.

In summary, the future of automated cyber threat attribution lies in integrating advanced AI techniques, developing real-time and explainable models, enhancing adversarial robustness, considering ethical and privacy implications, and benchmarking datasets and methods. Collaborative efforts within and across national borders will be essential for advancing the field and effectively combating the evolving landscape of cyber threats.

VIII. Conclusion

This survey provides an extensive review of the current landscape in cyber threat attribution and discusses the critical challenges and advancements that have shaped the field. The survey starts by exploring key artifacts that contribute to attribution and provides a taxonomy for scattered attribution artifacts. The taxonomy provides a standard way to understand and measure the possible attribution artifacts based on various qualitative measures. This survey also explores the available attribution datasets in the public domain and provides a classification of attribution datasets and their comparison. We further present the classification of available attribution methods based on the type of algorithm used and the dataset used for attribution. A comprehensive review of current literature on APT attribution is presented, along with critical comments and detailed discussions. Further, we analyze APT attribution's current state and mention challenges toward advancing attribution methodologies. These comprehensive analysis of current literature led us to understand the future research directions in this domain and present the open research problems that researchers and practitioners can look into to advance the current state of APT attribution. By synthesizing these insights, this survey provides a comprehensive overview of the current state of APT attribution and offers valuable guidance for future research. We believe this work will inspire new advancements and practical solutions in threat attribution, ultimately fortifying our defenses against sophisticated cyber threats.

Appendix
Artifact analysis reasoning

This section provides a logical analysis of the listed artifacts based on the selected measures.

TABLE VIII: Artifact analysis reasoning glimpse

Artifact Name	Relevance	Integrity	Credibility	Timeliness	Accessibility
Indicator of Compromise (IoCs)	Highly relevant as if IOCs like IP address, emails, cryptographic hashes are directly matched with any historical incident, then it may help in attribution confidently	It can be easily manipulated, spoofed, or reused by different threat actors, reducing their integrity in providing reliable attribution evidence	Often come from a mix of reliable open-source intelligence and private sector reports	Available shortly after or even during an attack, allowing for quick detection and response	Often easy to collect and analyze, as they are usually widely shared and well-documented
C&C Infrastructure	Highly relevant as it often directly links to specific threat actors' operations, providing a crucial link in identifying and attributing attacks.	It can be easy for attackers to disguise or alter without compromising their operations	Typically include in-depth network analysis and cybersecurity experts, making them highly credible and reliable for accurate attribution	Can be identified and analyzed relatively quickly, but often requires ongoing monitoring and analysis to be effective in real-time	Requires specific tools and expertise, and it can be challenging to detect active C&C servers
Crypto Currency Transaction	While they can be useful for tracking financial transactions, cryptocurrency transactions are often anonymized and lack direct relevance to specific threat actor attribution	While the blockchain provides transparency, the anonymity and potential use of mixers reduce their reliability for attribution	Often derived from blockchain analysis, which is transparent and traceable	Often takes time due to the need to analyze blockchain data, and relevant information may only become clear after the transactions have been fully processed	Tracking and analyzing cryptocurrency transactions can be complex due to the need for specialized tools and the anonymized nature of transactions
Tactics, Techniques & Procedures (TTPs)	Highly relevant as they represent the behavioral patterns of threat actors, which are often unique and difficult to fake, providing strong attribution evidence	High integrity because they represent consistent behavioral patterns that are hard for attackers to alter without changing their operational methods	Usually derived from detailed cybersecurity research and historical attack patterns, often published by reputable security organizations, ensuring high veracity	Generally identified over time through pattern analysis and research, and often emerge after multiple incidents	Accessible through threat intelligence reports and databases, but they require in-depth analysis and understanding to link them accurately to specific actors
Malware & Code Analysis	Analyzing malware and code can reveal unique signatures and methods that are often tied to specific threat groups, making this artifact highly relevant for attribution	High integrity because unique code signatures and methods are difficult to fake or alter significantly without detection	typically conducted by skilled cybersecurity researchers and forensic experts, ensuring that the sources are reliable and the findings are trustworthy	Typically time-consuming, requiring in-depth investigation that can delay actionable insights	Requires significant technical expertise and specialized tools to decompile and analyze code, making it less accessible to general analysts
Toolchains	The tools and software used by attackers can be indicative of specific groups, particularly when tied to known actor methodologies, making this highly relevant	It can be changed based on attacker group's preference and advancements in advanced tools	Often linked to direct observations by security professionals, including forensic investigations, providing high veracity	Requires detailed investigation, which can take time, but once identified, it can provide timely insights into ongoing campaigns	Identifying and analyzing a threat actor's toolchain requires detailed investigation and correlation of multiple artifacts, which can be resource-intensive
Language & Writing Style	Language and writing style can help identify specific individuals or groups, but they are less definitive on their own and may require corroboration with other artifacts	While unique, it can be imitated or altered, requiring careful consideration to avoid misattribution	Usually extracted from collected malware samples or received spam emails	Typically takes time, especially if it requires gathering sufficient samples for a conclusive analysis, leading to low timeliness	requires access to relevant samples and linguistic expertise, making it moderately accessible
Public Claims and Hacker Forums	While public claims and discussions on hacker forums can provide insights, they are not always reliable and may require validation, giving them medium relevance	Low integrity due to the potential for false claims, misinformation, or deliberate deception by actors	Often unverified and can be intentionally misleading, leading to a low level of veracity in these sources	Can emerge quickly and sometimes delay	relatively easy to access through open-source intelligence (OSINT), though its credibility varies
Threat Intelligence Databases	These databases compile various artifacts and intelligence but rely on the quality and context of the data, providing medium relevance in the attribution process	It varies depending on the source and verification processes, making them moderately reliable for attribution	Typically curated by established cybersecurity organizations, but sometimes may be updated by individual security researchers, which might raise concern regarding source integrity	Updated regularly, but there can be a delay between the detection of new threats and the inclusion of information in the database	These databases are designed to be accessible to analysts, providing structured information that is easy to query and analyze

Continue on the next page

TABLE VIII: Artifact analysis reasoning glimpse

Artifact Name	Relevance	Integrity	Credibility	Timeliness	Accessibility
Geopolitical	It can offer valuable insights into potential motivations and likely actors, but it needs to be combined with other evidence for precise attribution	Medium integrity, as it is interpretative and may be influenced by biases or incomplete information, but it still provides valuable context	Can be reliable, but they may also be influenced by political biases or incomplete information	Requires a broad and often slow-gathering of information, making it less timely for immediate threat attribution and response	Requires a good understanding of current geopolitical contexts and access to relevant intelligence reports
Victimology	Understanding the target's profile can be highly relevant, as specific threat actors often target certain industries or regions, providing clues to their identity	It relies on patterns that could potentially be aid if attackers intentionally target victims	Often based on reported incidents, patterns and data collected from victims	Can provide timely insights once sufficient data on targeted entities is gathered, but this often happens after multiple incidents have occurred	Requires detailed analysis of attack patterns and affected entities, which can be resource-intensive but is generally accessible with proper data
Post-Incident Communication	It can offer insights into the actors involved, but it is less direct and may require interpretation, giving it medium relevance	It can provide reliable insights, but may also be subject to manipulation or deception	Usually comes from victim's conversation with threat actor	typically emerge with a delay	This information can be accessible through monitoring communication channels, but it requires active intelligence gathering and analysis
Time Zone Analysis	While time zone analysis can suggest the likely origin of the threat actor, it is a circumstantial piece of evidence	It is circumstantial and can be easily misrepresented by attackers using tools to obfuscate their actual location	Often based on generally reliable metadata and logs	can be conducted quickly using metadata from logs or other artifacts, providing timely insights into the possible origin of the attack	Relatively straightforward and can often be conducted using metadata from logs and other time-stamped artifacts

REFERENCES

[1] A. Alshamrani, S. Myneni, A. Chowdhary, and D. Huang, "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1851–1877, 2019.

[2] N. Sun, M. Ding, J. Jiang, W. Xu, X. Mo, Y. Tai, and J. Zhang, "Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 3, pp. 1748–1774, 2023.

[3] S. Caltagirone, A. Pendergast, and C. Betz, "The diamond model of intrusion analysis," *Threat Connect*, vol. 298, no. 0704, pp. 1–61, 2013.

[4] S. Alrabaee, P. Shirani, M. Debbabi, and L. Wang, "On the feasibility of malware authorship attribution," in *Foundations and Practice of Security: 9th International Symposium, FPS 2016, Québec City, QC, Canada, October 24-25, 2016, Revised Selected Papers 9*, 2017, pp. 256–272.

[5] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, and Z. Tian, "CSKG4APT: A cybersecurity knowledge graph for advanced persistent threat organization attribution," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[6] A. Warikoo, "The triangle model for cyber threat attribution," *Journal of Cyber Security Technology*, vol. 5, no. 3-4, pp. 191–208, 2021.

[7] T. Steffens, *Attribution of Advanced Persistent Threats*, Springer, 2020.

[8] U. Noor, Z. Anwar, T. Amjad, et al., "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise," *Future Generation Computer Systems*, vol. 96, pp. 227–242, 2019.

[9] D. Schlette, M. Caselli, and G. Pernul, "A comparative study on cyber threat intelligence: The security incident response perspective," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2525–2556, 2021.

[10] Egloff, Florian J. "Public attribution of cyber intrusions." *Journal of cybersecurity*, vol. 6, no. 1, 2020, tyaa012. Oxford University Press.

[11] Shenderovitz, Gil, and Nir Nissim. "Bon-APT: Detection, attribution, and explainability of APT malware using temporal segmentation of API calls." *Computers & Security*, vol. 142, 2024, p. 103862. Elsevier.

[12] Ahmad, A., Webb, J., Desouza, K. C., & Boorman, J. (2019). Strategically-motivated advanced persistent threat: Definition, process, tactics and a disinformation model of counterattack. Computers & Security, 86, 402–418. Elsevier.

[13] MISP Project. (n.d.). MISP Galaxy Clusters. Retrieved July 18, 2024, from https://www.misp-project.org/tags/Galaxy/

[14] Na Xu, Shudong Li, Xiaobo Wu, Weihong Han, and Xiaojing Luo, "An APT Malware Classification Method Based on Adaboost Feature

Selection and LightGBM," in *2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*, IEEE, 2021, pp. 635–639.

[15] P. R. Brandao, "Advanced Persistent Threats (APT)-Attribution-MICTIC Framework Extension," *J. Comput. Sci*, vol. 17, pp. 470–479, 2021.

[16] VirusTotal. (n.d.). VirusTotal - Home Upload. Retrieved July 18, 2024, from https://www.virustotal.com/gui/home/upload

[17] T. Rid and B. Buchanan, "Attributing cyber attacks," *Journal of Strategic Studies*, vol. 38, no. 1-2, pp. 4–37, 2015.

[18] Marquis-Boire, M., Marschalek, M., & Guarnieri, C. (2015). Big game hunting: The peculiarities in nation-state malware research. *Black Hat, Las Vegas, NV, USA*.

[19] "APTnotes," GitHub repository, 2015. [Online]. Available: https://github.com/aptnotes/data. [Accessed: June 29, 2024].

[20] M. Tatam, B. Shanmugam, S. Azam, and K. Kannoorpatti, "A review of threat modelling approaches for APT-style attacks," *Heliyon*, vol. 7, no. 1, 2021.

[21] Bartholomew, Brian, and Juan Andres Guerrero-Saade. "Wave your false flags! deception tactics muddying attribution in targeted attacks." *Virus Bulletin Conference*, 2016, pp. 1-9.

[22] Q. Wang, H. Yan, and Z. Han, "Explainable APT Attribution for Malware Using NLP Techniques," in *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, pp. 70–80, 2021, IEEE.

[23] MITRE ATT&CK. (n.d.). APT29. Retrieved July 18, 2024, from https://attack.mitre.org/groups/G0016/

[24] F. Skopik and T. Pahi, "Under false flag: using technical artifacts for cyber attack attribution," *Cybersecurity*, vol. 3, pp. 1–20, 2020, Springer.

[25] United States Department of Justice, "U.S. Charges Five Chinese Military Hackers For Cyber Espionage Against U.S. Corporations And A Labor Organization For Commercial Advantage", 2014, [Online]. https://www.fbi.gov/contact-us/field-offices/pittsburgh/news/press-releases/u.s.-charges-five-chinese-military-hackers-with-cyber-espionage-against-u.s.-corporations-and-a-labor-organization-for-commercial-advantage

[26] G. Laurenza and R. Lazzeretti, "dAPTaset: A comprehensive mapping of APT-related data," in *Computer Security: ESORICS 2019 International Workshops, IOSec, MSTEC, and FINSEC, Luxembourg City, Luxembourg, September 26–27, 2019, Revised Selected Papers 2*, Springer, 2020, pp. 217–225.

[27] VirusShare. (n.d.). VirusShare. Retrieved July 11, 2024, from https://virusshare.com

[28] Sedding, Julian, and Dimitar Kazakov. "Wordnet-based text document clustering." In *Proceedings of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND 2004)*, 2004, pp. 104-113.

[29] M. Kida and O. Olukoya, "Nation-State Threat Actor Attribution Using Fuzzy Hashing," *IEEE Access*, vol. 11, pp. 1148–1165, 2022, IEEE.

[30] "MalwareBazaar", ABUSE—ch, [Online]. https://bazaar.abuse.ch

[31] C. Boot, "Applying Supervised Learning on Malware Authorship Attribution," Master thesis, Radboud University, Institute for Computing and Information , 2019.

[32] Cyber Monitor, "APT CyberCriminal Campaign Collections," GitHub repository, 2024. [Online]. Available: https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections. [Accessed: April 29, 2024].

[33] Patsakis, C., Arroyo, D., & Casino, F. (2024). The Malware as a Service ecosystem. *arXiv preprint arXiv:2405.04109*.

[34] MalShare. (n.d.). MalShare. Retrieved July 13, 2024, from https://malshare.com

[35] Irshad, E., & Siddiqui, A. B. (2024). Context-aware cyber-threat attribution based on hybrid features. ICT Express. Elsevier.

[36] Herbert Lin, "Attribution of malicious cyber incidents: From soup to nuts," *Journal of International Affairs*, vol. 70, no. 1, pp. 75–137, 2016. [Online]. Available: JSTOR.

[37] Wikipedia. (n.d.). Russo-Ukrainian Cyberwarfare. Retrieved July 18, 2024, from https://en.wikipedia.org/wiki/Russo-Ukrainian_cyberwarfare

[38] Selva Birunda, S., and R. Kanniga Devi. "A review on word embedding techniques for text classification." *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, 2021, pp. 267-281. Springer.

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

[40] K. Edie, C. Mckee, and A. Duby, "Extending Threat Playbooks for Cyber Threat Intelligence: A Novel Approach for APT Attribution," in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1-6, 2023.

[41] "Hatching Triage", Hatching, [Online]. https://hatching.io/triage/

[42] Jason Gray, Daniele Sgandurra, Lorenzo Cavallaro, and Jorge Blasco. Identifying Authorship in Malicious Binaries: Features, Challenges & Datasets. *ACM Computing Surveys*, ACM New York, NY, 2024.

[43] Sharma, A., Gupta, B. B., Singh, A. K., & Saraswat, V. K. (2023). Advanced persistent threats (APT): Evolution, anatomy, attribution and countermeasures. Journal of Ambient Intelligence and Humanized Computing, 14(7), 9355-9381. Springer.

[44] FireEye, "Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims With SUNBURST Backdoor", FireEye (Now part of Google Cloud), 2020, [Online]. https://cloud.google.com/blog/topics/threat-intelligence/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor/

[45] Hwang, S., & Kim, T. S. (2023). An Exploratory Study on Artifacts for Cyber Attack Attribution Considering False Flag: Using Delphi and AHP methods. IEEE Access.

[46] Daszczyszak, R., Ellis, D., Luke, S., and Whitley, S. (2019). TTP-based hunting. *MITRE CORP MCLEAN VA, Tech. Rep*.

[47] Malpedia. (n.d.). Malpedia. Retrieved July 15, 2024, from https://malpedia.caad.fkie.fraunhofer.de

[48] Lior Perry, Bracha Shapira, and Rami Puzis. No-doubt: Attack attribution based on threat intelligence reports. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 80–85. IEEE, 2019.

[49] Stoll, Cliff. *The cuckoo's egg: tracking a spy through the maze of computer espionage*. Simon and Schuster, 2005.

[50] "Hybrid Analysis", Hybrid Analysis, [Online]. https://hybrid-analysis.com

[51] Tsagourias, N., & Farrell, M. (2020). Cyber attribution: technical and legal approaches and challenges. *European Journal of International Law*, 31(3), 941–967. Oxford University Press UK.

[52] Flashpoint Intel Team,"Evolving Tactics: How Russian APT Groups Are Shaping Cyber Threats in 2024," Flashpoint, 2024, [Online]. Available: https://flashpoint.io/blog/russian-apt-groups-cyber-threats/

[53] Qinqin Wang, Hanbing Yan, Chang Zhao, Rui Mei, Zhihui Han, and Yu Zhou. *APT attribution for malware based on time series shapelets*. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 769–777, 2022. IEEE.

[54] Boebert, W. E. (2010). A survey of challenges in attribution. In *Proceedings of a Workshop on Deterring CyberAttacks* (pp. 41–54).

[55] "paperAPTvsMalware", Github Repository, [Online]. https://github.com/lgmanzan/paperAPTvsMalware/tree/main

[56] S. Quintero-Bonilla and A. M. del Rey, "A new proposal on the advanced persistent threat: A survey," *Applied Sciences*, vol. 10, no. 11, p. 3874, 2020.

[57] Center for Strategic and International Studies. (2022). Cyber Operations During the Russo-Ukrainian War. Retrieved July 18, 2024, from https://www.csis.org/analysis/cyber-operations-during-russo-ukrainian-war

[58] Kelsie Edie, Cole McKee, and Adam Duby, *Extending Threat Playbooks for Cyber Threat Intelligence: A Novel Approach for APT Attribution*, in *2023 IEEE 11th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, 2023.

[59] "APTMalware", Github Repository, [Online]. https://github.com/cyber-research/APTMalware

[60] Microsoft Threat Intelligence, "Analyzing Solorigate, the compromised DLL file that started a sophisticated cyberattack, and how Microsoft Defender helps protect customers", Microsoft, 2020, [Online]. https://www.microsoft.com/en-us/security/blog/2020/12/18/analyzing-solorigate-the-compromised-dll-file-that-started-a-sophisticated-cyberattack-and-how-microsoft-defender-helps-protect/

[61] Robert K. Knake, "Untangling attribution: Moving to accountability in cyberspace," *Prepared Statement before the Subcommittee on Technology and Innovation, Committee on Science and Technology, Hearing: Planning for the Future of Cyber Attack*, 2010.

[62] T. D. Wagner, K. Mahbub, E. Palomar, and A. E. Abdallah, "Cyber threat intelligence sharing: Survey and research directions," *Computers & Security*, vol. 87, p. 101589, 2019.

[63] S3Lab, Information Security Group, Royal Holloway, University of London, *APT Classification Project*, Available at: https://s3lab.isg.rhul.ac.uk/project/aptclass/, Accessed: [June 29, 2024].

[64] Goel, S., & Nussbaum, B. (2021). Attribution across cyber attack types: network intrusions and information operations. IEEE Open Journal of the Communications Society, 2, 1082-1093. IEEE.

[65] Congressional Research Service. Ransomware: Analysis and Recommendations for Congress. https://crsreports.congress.gov/product/pdf/R/R46974/4. Accessed: 2024-08-03.

[66] Y. Xu, Y. Fang, C. Huang, and Z. Liu, "Hghan: hacker group identification based on heterogeneous graph attention network," *Information Sciences*, vol. 612, pp. 848–863, 2022, Elsevier.

[67] Mandiant, *Homepage*, Available at: https://www.mandiant.com, Accessed: [June 29, 2024].

[68] Swate, C., Sithungu, S., & Lebea, K. (2024). An Analysis of Cyber-warfare Attribution Techniques and Challenges. In *Proceedings of the European Conference on Cyber Warfare and Security*, 23(1), 552–559.

[69] "Advanced Persistent Threat (APT) Malware Dataset – 2020", Cyber Science Lab, 2020, [Online] https://cybersciencelab.com/advanced-persistent-threat-apt-malware-dataset/

[70] Umara Noor, Sawera Shahid, Rimsha Kanwal, and Zahid Rashid. A Machine Learning based Empirical Evaluation of Cyber Threat Actors High Level Attack Patterns over Low level Attack Patterns in Attributing Attacks. *arXiv preprint arXiv:2307.10252*, 2023.

[71] Unit 42, Palo Alto Networks, *Homepage*, Available at: https://unit42.paloaltonetworks.com, Accessed: [June 25, 2024].

[72] Sangwoo Lee and Jungwon Cho, "Malware Authorship Attribution Model using Runtime Modules based on Automated Analysis," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 1-2, pp. 214–220, 2022.

[73] Trellix, *Homepage*, Available at: https://www.trellix.com/en-in/, Accessed: [Jul. 11, 2024].

[74] L. Qiang, Y. Zeming, L. Baoxu, J. Zhengwei, and Y. Jian, "Framework of cyber attack attribution based on threat intelligence," in *Interoperability, Safety and Security in IoT: Second International Conference, InterIoT 2016 and Third International Conference, SaSeIoT 2016, Paris, France, October 26-27, 2016, Revised Selected Papers 2*, 2017, pp. 92–103.

[75] "Extending-Threat-Playbooks-for-APT-Attribution", Github Repository, [Online]. https://github.com/KelsieEdie/Extending-Threat-Playbooks-for-APT-Attribution

[76] W. Detlefsen, "Cyber attacks, attribution, and deterrence: Three case studies," *Leaventh: US Army Command and General Staff College*, 2015.

[77] T. Pahi and F. Skopik, "Cyber attribution 2.0: Capture the false flag," in *Proceedings of the 18th European Conference on Cyber Warfare and Security (ECCWS 2019)*, 2019, pp. 338–345.

[78] CrowdStrike, *Homepage*, Available at: https://www.crowdstrike.com/en-us/, Accessed: [Jul. 07, 2024].

[79] Ishai Rosenberg, Guillaume Sicard, and Eli David. DeepAPT: nation-state APT attribution using end-to-end deep neural networks. In *Artificial Neural Networks and Machine Learning–ICANN 2017: 26th International Conference on Artificial Neural Networks, Alghero, Italy, September 11-14, 2017, Proceedings, Part II*, volume 26, pages 91–99. Springer, 2017.

[80] "MITRE ATT&CK® Navigator", The MITRE Corporation, [Online]. https://mitre-attack.github.io/attack-navigator/

[81] Rani, Nanda, Bikash Saha, Vikas Maurya, and Sandeep Kumar Shukla. TTPHunter: automated extraction of actionable intelligence as TTPs from narrative threat reports. In *Proceedings of the 2023 Australasian Computer Science Week*, pages 126–134, 2023.

[82] Lockheed Martin, "The Cyber Kill Chain", Lockheed Martin, 2011, [online]. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[83] Nan Xiao, Bo Lang, Ting Wang, and Yikai Chen. APT-MMF: An advanced persistent threat actor attribution method based on multimodal and multilevel feature fusion. *Computers & Security*, page 103960, 2024.

[84] "Cyber-Attack-Patterns-Dataset", GitHub Repository, [Online]. https://github.com/UmaraNoor/Cyber-Attack-Patterns-Dataset

[85] Shudong Li, Qianqing Zhang, Xiaobo Wu, Weihong Han, and Zhihong Tian, "Attribution classification method of APT malware in IoT using machine learning techniques," *Security and Communication Networks*, vol. 2021, pp. 1–12, 2021, Hindawi Limited.

[86] MITRE ATT&CK. (n.d.). APT32. Retrieved July 18, 2024, from https://attack.mitre.org/groups/G0032/

[87] Krebs, Brian. "Malicious Domain in SolarWinds Hack Turned into Killswitch." *Krebs on Security*, December 2020. Retrieved from https://krebsonsecurity.com/2020/12/malicious-domain-in-solarwinds-hack-turned-into-killswitch/.

[88] Y. Mei, W. Han, S. Li, X. Wu, K. Lin, and Y. Qi, "A Review of Attribution Technical for APT Attacks," in *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pp. 512–518, 2022.

[89] Cyber Conflicts, "Case Study Viasat", Cyber Peach Instittue, 2022, [Online]. https://cyberconflicts.cyberpeaceinstitute.org/law-and-policy/cases/viasat#attribution

[90] Husari, Ghaith, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of CTI sources. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 103–115, 2017.

[91] "dAPTaset", GitHub Repository, [Online]. https://github.com/GiuseppeLaurenza/dAPTaset

[92] J. Li, J. Liu, and R. Zhang, "Advanced Persistent Threat Group Correlation Analysis via Attack Behavior Patterns and Rough Sets," *Electronics*, vol. 13, no. 6, p. 1106, 2024.

[93] F-Secure, *Homepage*, Available at: https://www.f-secure.com/en, Accessed: [Jul. 11, 2024].

[94] Yangyang Mei, Weihong Han, Shudong Li, Kaihan Lin, and Cui Luo, "A hybrid intelligent approach to attribute Advanced Persistent Threat Organization using PSO-MSVM Algorithm," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 4262–4272, 2022, IEEE.

[95] Flare Research Team. Telegram and OSINT Investigations: An Essential Platform in 2023. https://flare.io/learn/resources/blog/telegram-investigation/. Accessed: 2024-08-03.

[96] Schnirch, Lilli P.M. Exploring Influencing Behaviours in Ransomware Negotiations: Comparing Parties and Identifying Patterns. B.S. thesis, University of Twente, 2024.

[97] , Suleyman Ozarslan, "Tactics, Techniques, and Procedures (TTPs) Used in the SolarWinds Breach", Picus Security. 2020. [Online]. https://www.picussecurity.com/resource/blog/ttps-used-in-the-solarwinds-breach

[98] Tang, B., Wang, J., Yu, Z., Chen, B., Ge, W., Yu, J., & Lu, T. (2022). Advanced Persistent Threat intelligent profiling technique: A survey. Computers and Electrical Engineering, 103, 108261. Elsevier.

[99] Federal Bureau of Investigation. (n.d.). Park Jin Hyok. Retrieved July 18, 2024, from https://www.fbi.gov/wanted/cyber/park-jin-hyok

[100] Rani, Nanda, Bikash Saha, Vikas Maurya, and Sandeep Kumar Shukla. TTPXHunter: Actionable Threat Intelligence Extraction as TTPs from Finished Cyber Threat Reports. *arXiv preprint arXiv:2403.03267*, 2024.

[101] ANTONY J. BLINKEN, "Attribution of Russia's Malicious Cyber Activity Against Ukraine", U.S. Department of State, 2022 [Online] https://www.state.gov/attribution-of-russias-malicious-cyber-activity-against-ukraine/

[102] Yinghai Zhou, Yi Tang, Ming Yi, Chuanyu Xi, Hai Lu. CTI view: APT threat intelligence analysis system. *Security and Communication Networks*, 2022;2022:1–15. Hindawi Limited.

[103] "APT Groups and Operations", Google Sheet, [Online]. https://docs.google.com/spreadsheets/d/1H9_xaxQHpWaa4O_Son4Gx0YOIzlcBWMsdvePFX68EKU

[104] N. Pitropakis, E. Panaousis, A. Giannakoulias, G. Kalpakis, R. D. Rodriguez, and P. Sarigiannidis, "An enhanced cyber attack attribution framework," in *Trust, Privacy and Security in Digital Business: 15th International Conference, TrustBus 2018, Regensburg, Germany, September 5–6, 2018, Proceedings 15*, 2018, pp. 213–228.

[105] The Evolution of Cyber Attribution. https://www.american.edu/sis/centers/security-technology/the-evolution-of-cyber-attribution.cfm. Accessed: 2024-08-03.

[106] "Group", The MITRE Corporation, [Online]. https://attack.mitre.org/groups/

[107] S. Naveen, Rami Puzis, and Kumaresan Angappan. Deep learning for threat actor attribution from threat reports. In *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–6. IEEE, 2020.

[108] G. Ahmadi-Assalemi, H. M. Al-Khateeb, G. Epiphaniou, J. Cosson, H. Jahankhani, and P. Pillai, "Federated blockchain-based tracking and liability attribution framework for employees and cyber-physical objects in a smart workplace," in *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*, 2019, pp. 1–9.

[109] Kaspersky, "Threat Attribution Engine Whitepaper," *Kaspersky*, [Online]. Available: https://media.kaspersky.com/en/business-security/enterprise/threat-attribution-engine-whitepaper.pdf. [Accessed: Jul. 11, 2024].

[110] Hawdon, James. Cybercrime: Victimization, perpetration, and techniques. *American Journal of Criminal Justice*, 46(6):837–842, 2021.

[111] "Software", The MITRE Corporation, [Online]. https://attack.mitre.org/software/

[112] M. R. Korver, C. A. Pelker, and E. Poteat, "Attribution in cryptocurrency cases," *Dep't of Justice Journal of Federal Law and Practice*, vol. 67, p. 233, 2019.

[113] Christopher Bing, "Suspected Russian hackers spied on U.S. Treasury emails", Reuters, 2020, [Online]. https://www.reuters.com/article/BigStory12/idUSKBN28N0PG/

[114] Y. Wu, C. Huang, X. Zhang, and H. Zhou, "GroupTracer: Automatic attacker TTP profile extraction and group cluster in Internet of things," *Security and Communication Networks*, vol. 2020, pp. 1–14, 2020, Hindawi Limited.

[115] "Campaigns", The MITRE Corporation, [Online]. https://attack.mitre.org/campaigns/

[116] Hamed Haddadpajouh, Amin Azmoodeh, Ali Dehghantanha, and Reza M Parizi. MVFCC: A multi-view fuzzy consensus clustering model for malware threat attribution. *IEEE Access*, 8:139188–139198, 2020.

[117] F. Jaafar, F. Avellaneda, and E.-H. Alikacem, "Demystifying the cyber attribution: An exploratory study," in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2020, pp. 35–40.

[118] MITRE, "ATT&CK Framework", The MITRE Corporation, [Online]. https://attack.mitre.org.

[119] Vinay Sachidananda, Rajendra Patil, Akshay Sachdeva, Kwok-Yan Lam, and Liu Yang. APTer: Towards the Investigation of APT Attribution. In *2023 IEEE Conference on Dependable and Secure Computing (DSC)*, pages 1–10. IEEE, 2023.

[120] Kaspersky, *Homepage*, Available at: https://www.kaspersky.co.in, Accessed: [Jul. 10, 2024].

[121] F. Avellaneda, E.-H. Alikacem, and F. Jaafar, "Using Attack Pattern for Cyber Attack Attribution," in *2019 International Conference on Cybersecurity (ICoCSec)*, 2019, pp. 1–6.

[122] Federal Bureau of Investigation. (n.d.). Kim Il. Retrieved July 18, 2024, from https://www.fbi.gov/wanted/cyber/kim-il

[123] Peter W. Singer and Allan Friedman, *Cybersecurity: What everyone needs to know*, OUP USA, 2014.

[124] Saha, Bikash, Nanda Rani, and Sandeep Kumar Shukla. MalXCap: A Method for Malware Capability Extraction. In *International Conference on Information Security Practice and Experience*, pages 230–249. Springer, 2023.

[125] Kaspersky Lab. "Sunburst backdoor - Kazuar." *Securelist*, 13 January 2021. Available at: https://securelist.com/sunburst-backdoor-kazuar/99981/.

[126] V. Cherniei, S. Cherniavskyi, V. Babanina, and O. Tykho, "Criminal liability for cryptocurrency transactions: Global experience," *European Journal of Sustainable Development*, vol. 10, no. 4, p. 304, 2021.

[127] Blue Goat Cyber. (n.d.). The Challenge of Attribution in Cyber Attacks. Retrieved July 18, 2024, from https://bluegoatcyber.com/blog/the-challenge-of-attribution-in-cyber-attacks/

[128] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.s

[129] "Campaign C0024. SolarWinds Compromise" Available at: https://attack.mitre.org/campaigns/C0024/. Accessed: [Jul. 11, 2024].

[130] F. J. Egloff and M. Smeets, "Publicly attributing cyber attacks: a framework," *Journal of Strategic Studies*, vol. 46, no. 3, pp. 502–533, 2023.

[131] K. Kim, Y. Shin, J. Lee, and K. Lee, "Automatically attributing mobile threat actors by vectorized ATT&CK matrix and paired indicator," *Sensors*, vol. 21, no. 19, p. 6522, 2021.

[132] Cyberint Research Team. Behind the Mask of Anonymous Sudan: An Analysis. https://cyberint.com/blog/research/anonymous-sudan-an-analysis/. Accessed: 2024-08-03.

[133] Y. Shin, K. Kim, J. J. Lee, and K. Lee, "ART: Automated Reclassification for Threat Actors Based on ATT&CK Matrix Similarity," in *2021 World Automation Congress (WAC)*, 2021, pp. 15–20.

[134] Federal Bureau of Investigation. (n.d.). Jon Chang Hyok. Retrieved July 18, 2024, from https://www.fbi.gov/wanted/cyber/jon-chang-hyok

[135] Aditya Grover, Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

[136] University of Washington, "CISA: Attack on SolarWinds "poses grave risk"", University of Washington, 2020 [Online]. https://ciso.uw.edu/2020/12/19/cisa-attack-on-solarwinds-poses-grave-risk/

[137] Alrawi, Omar, Moses Ike, Matthew Pruett, Ranjita Pai Kasturi, Srimanta Barua, Taleb Hirani, Brennan Hill, and Brendan Saltaformaggio. Forecasting malware capabilities from cyber attack memory images. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3523–3540, 2021.