IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 September 2024, accepted 28 September 2024, date of publication 3 October 2024, date of current version 14 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3473021

APPLIED RESEARCH

APTracker: A Comprehensive and Analytical Malware Dataset Based on Attribution to APT Groups

MOHAMAD ERFAN MAZAHERI AND ALIREZA SHAMELI-SENDI
Faculty of Computer Science and Engineering, Shahid Beheshti University (SBU), Tehran 19839 69411, Iran
Corresponding author: Alireza Shameli-Sendi (a_shameli@sbu.ac.ir)

ABSTRACT Malware poses a significant threat to organizations, necessitating robust countermeasures. One such measure involves attributing malware to its respective Advanced Persistent Threat (APT) group, which serves several purposes, two of the most important ones are: aiding in incident response and facilitating legal recourse. Recent years have witnessed a surge in research efforts aimed at refining methods for attributing malware to specific threat groups. These endeavors have leveraged a variety of machine learning and deep learning techniques, alongside diverse features extracted from malware binary files, to develop attribution systems. Despite these advancements, the field continues to beckon further investigation to enhance attribution methodologies. The basis of developing an effective attribution systems is to benefit from a rich dataset. Previous studies in this domain have meticulously detailed the process of model training and evaluation using distinct datasets, each characterized by unique strengths, weaknesses, and varying number of samples. In this paper, we scrutinize previous datasets from several perspectives while focusing on analyzing our dataset, which we claim is the most comprehensive in the realm of malware attribution. This dataset encompasses 64,440 malware samples attributed to 22 APT groups and spans a minimum of 40 malware families. The samples in the dataset span the years 2020 to 2024, and their developer APT groups originate from Russia, South Korea, China, USA, Nigeria, North Korea, Pakistan and Belarus. Its richness and breadth render it invaluable for future research endeavors in the field of malware attribution.

INDEX TERMS APT, attribution, dataset, malware attribution.

I. INTRODUCTION

Malware is any program developed with the intention of harming a computer, server, or network. Malware pursues different goals, and its features and capabilities are considered based on these goals. Some types of malware include viruses, worms, Trojans, ransomware, wipers, spyware, and Remote Access Trojans (RATs). The level of destruction caused by a malware can be very different, depending on its features and purpose. This variation in destruction includes both the number of victims and the amount of destruction and cost per victim [1]. Cyber threat actors encompass different types. These actors, who are actually individuals or groups that endanger cyber security, are classified based on their motivation, type of attack, and intended goal. Among these, we can mention script-kiddies, cybercriminals, hacktivists, and APTs, and so on [2]. Malwares developed by APT groups can be considered as the most sophisticated malwares.

The term ''APT'' was first used in 2006 [3]. An Advanced Persistent Threat (APT) refers to a highly planned and usually long lasting cyberattack that often run multiple long lasting campaigns. Usually, these attacks focus on specific organizations or companies, trying to break into their computer systems. They mainly use malwares, and the groups behind these attacks are called APT groups [4].

Attribution, a pivotal process utilized across various disciplines, holds particular significance in computer engineering. Within this context, attribution entails identifying the

The associate editor coordinating the review of this manuscript and approving it for publication was Leandros Maglaras.

developer or developer group behind a software, based on a set of discernible features [5], [6], [7], [8]. This attribution process unfolds at two distinct levels: source code attribution [6], [9], [10], [11], [12] and binary file attribution [4], [13], [14], [15], [16], [17]. Source code attribution is pertinent when access to high-level programming code (e.g., C, C++, Golang) is available, whereas binary file attribution involves analyzing the compiled version of the code, presented in assembly language.

In essence, software attribution involves scrutinizing an array of features, encompassing stylistic, lexical, syntactic, semantic, behavioral, and application-specific attributes at the source code level. Conversely, at the binary level, features such as strings, implementational nuances, infrastructural aspects, source code-related characteristics corresponding to the binary, and assembly language features are examined.

Attributing malware to APT groups serves as a pivotal step in elucidating the overarching objectives of threat actors, enabling effective incidence response strategies and legal recourse. Malware attribution, in essence, falls under the broader umbrella of software attribution. In the context of this paper, malware attribution pertains specifically to attributing malware to APT groups. APT groups exhibit distinct patterns in their attack methodologies across various incidents, generating identifiable features that attribution systems leverage to discern the originating malware developer group [4].

The expediency of malware analysis and subsequent identification of the APT group directly correlates with the promptness of understanding and responding to the attack. While attribution systems have demonstrated promising results in attributing malwares, the future trajectory of research in this domain necessitates extensive endeavors from both defensive and offensive standpoints.

In recent years, the proliferation of machine learning and deep learning methodologies within the realm of security has witnessed researchers extensively leveraging these approaches for malware attribution, considering a myriad of features in the process. Furthermore, commendable efforts have been directed towards compiling datasets conducive to the attribution process.

Central to the infrastructure of malware attribution systems lies the dataset employed for training and evaluating machine learning methodologies. These datasets comprise exemplars of malware categorized into distinct APT groups, often sourced from technical reports or sandbox-generated reports in the majority of research endeavors.

Overall, the paper makes the following contributions:

- Introducing our dataset: ‘‘APTracker’’, the most comprehensive malware dataset structured around APT groups, enriching the research landscape with unparalleled depth and breadth of data. The dataset is comprehensive in that it includes a large number of malware samples representing as many APT groups and malware families as possible. Additionally, a sufficient number of samples has been included from each malware family.
- Offering an analytical perspective on the diverse malware families encompassed within APTracker, shedding light on their intricacies and implications.
- Conducting a meticulous review and evaluation of existing datasets within the realm of malware attribution, accompanied by a thorough exploration of the challenges inherent in dataset preparation, thus providing valuable insights for future research endeavors in this domain.

The reminder of this paper is organized as follows. Section II delineates the distinctive features of datasets utilized in malware attribution and elucidates their divergence from other malware datasets, culminating in a comprehensive review and evaluation of extant datasets. Section III unveils the focal point of this paper: APTracker, the most comprehensive repository in the field of malware attribution, meticulously curated as the primary contribution of this research endeavor. By grappling with the challenges and constraints outlined earlier, APTracker emerges as a significant stride in advancing the domain of malware attribution. Concluding insights drawn from the various themes expounded upon throughout the paper are synthesized in Section IV, encapsulating the overarching findings and implications elucidated within this discourse.

II. EVALUATING PRESENTED DATASETS

A. DATASET FEATURES IN MALWARE ATTRIBUTION

The datasets utilized in malware attribution systems are comprised of malware samples classified according to APT groups. Typically, reports detailing these malwares serve as input for training machine learning models, offering a rich source of information encompassing diverse aspects, from static to dynamic features. The scope of these features is very diverse. Examples include imported DLLs, different sections of the malware file, strings, malware-related processes, network features, dropped files, registry keys activities, and system calls, etc. These reports may stem from technical reports of cyber attacks and malwares analyzed by digital forensics and incident response teams [19], [20], [21], or from sandbox reports, such as those generated by[1] [22] upon execution of malware samples.

It is essential to distinguish the dataset utilized in attribution processes from other datasets pertaining to different facets of malware, such as malware detection. In attribution datasets, segregating malware samples according to APT groups is imperative. Ideally, each APT group encompasses various malware families. This segmentation enables the identification of the APT group responsible for a particular malware, based on patterns discerned from attribution systems, thereby facilitating prompt recognition and response to emerging threats.

[1]Malware analysis in the Cuckoo sandbox can be conducted both through the infrastructure and the web user interface provided by the https://cuckoo.cert.ee/ URL, as well as locally using virtual machines

B. EXAMINATION OF PROVIDED DATASETS

Prior to delving into the discussion, it is pertinent to clarify that this paper primarily focuses on reviewing datasets presented in various research endeavors, without delving extensively into the multifaceted dimensions of each study. However, providing a succinct overview of the machine learning models and features employed in the attribution systems is essential for readers to assess the compatibility of a dataset with specific research objectives.

The rationale behind briefly outlining these methodologies is rooted in the fact that the suitability of a dataset may vary depending on the machine learning model utilized. While an paper may demonstrate satisfactory performance metrics using a particular dataset, its compatibility with other research endeavors may not be guaranteed. Therefore, understanding the nuances of these methodologies aids in discerning the applicability of a dataset to diverse research contexts.

In an attempt to furnish readers with comprehensive insights, meticulous attention is devoted to delineating the characteristics of the presented datasets. This includes elucidating the number of APT groups covered, the distribution of malware samples across these groups, and the diversity of malware families represented within each group.

For instance, in a study by [3] in 2017, a deep neural network-based method for malware attribution was proposed. However, the dataset employed comprised 1,000 malware samples from Chinese and Russian APT groups, with limited details provided regarding the specific malware families utilized. This implies that a correct assessment of the dataset's balancedness cannot be expected. It is noteworthy that the aforementioned dataset remains unpublished, and no methodology has been provided to validate the accuracy of attributing malware to an APT group.

Similarly, [23], in 2018, leveraged a feature set processed using binary and TF-IDF vectors for attribution. The output of these two vectors is fed into the following classifiers: Decision Tree, Tree Bagging, Random Forest, Naïve Bayes, and Support Vector Machine. This dataset consisted of 1,088 malware samples from online repositories, distributed across 7 APT groups. Notably, the dataset was not publicly accessible, and no validation method was provided. The group with the highest number of samples comprises 434 malware samples, while the group with the lowest number of samples has 24 malware samples.

A noteworthy contribution in this domain is [24], published in 2019, which meticulously discussed the dataset utilized for malware attribution and subsequently made it available on GitHub. This dataset encompasses 3,594 malware samples attributed to 12 APT groups spanning 5 countries. Malware samples were collected for this work by leveraging open-source threat intelligence reports from various vendors, followed by downloads from VirusTotal. However, the method for labeling malware and the validation mechanism remain unspecified, while certain APT groups lack clarity regarding their associated malware families.

In 2019, [19] proposed a malware attribution approach based on high-level patterns extracted from technical reports, focusing on tactics, techniques, and procedures (TTPs). Five machine learning models—Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, and Deep Neural Network—were employed for malware attribution. The dataset utilized in this study comprised 327 technical reports sourced from public records documenting cyber incidents spanning from 2012 to 2018. While the dataset itself was not made publicly available, the selection of technical reports from publicly accessible sources serves as a validation method for its credibility.

Similarly, [20], also in 2019, introduced a methodology for malware attribution leveraging Natural Language Processing (NLP) techniques applied to malware technical reports. The dataset preparation process involved utilizing resources such as MITRE ATT&CK and APT Groups and Operations.[2] During data collection, reports lacking attribution or attributing attacks to multiple APT groups were excluded. Prominent APT groups with a substantial number of reports were prioritized, with additional reports sought through an automated tool using APT group aliases, followed by manual validation by experts. The resulting dataset comprised 249 technical reports documenting attacks attributed to 12 APT groups: APT28, Lazarus, Turla, OilRig, APT17, FIN7, APT3, APT29, menuPass, Rocket Kitten, Winnti, and Deep Panda. The distribution of reports varied, with APT28 featuring the highest count at 55, while Deep Panda had the fewest reports, totaling 10.

In 2020, [22] introduced a Multi-view Fuzzy Consensus Clustering model for attributing malware to their respective developer groups. Initially, the paper prepared reports on malware samples in the environment. The Decision Tree method was employed for training the dataset, comprising 1,200 malware samples from 5 APT groups: APT1, APT3, APT28, APT33, and APT37. Although the paper did not specify a validation method, the extraction of Cuckoo reports from malware samples and the utilization of their information suggest a means of validating the accuracy of classification. However, this dataset has not been made publicly available.

In 2020, [25] highlighted the importance of considering one class for each APT group and a separate class for non-APT group malware when training a multi-class classifier. The paper proposed using Isolation Forest for each class due to the heterogeneous nature of the last class and the imbalance in sample numbers. The dataset was sourced from the publicly released dAPTaset dataset, containing 9,000 unique malware hashes. By selecting Windows malwares from this dataset and expanding it using public sources, the paper obtained 2,086 malware samples attributed to 15 APT groups. These groups include APT28, APT29, APT30, Carbanak, Desert Falcon, Hurricane Panda, Lazarus Group, Mirage, Patchwork, Shiqiang, Transparent Tribe, Violin Panda, Volatile Cedar, and Winnti Group. The Patchwork

[2]https://airtable.com/shr3Po3DsZUQZY4we

group had the highest number of malware samples (559), while the Violin Panda group had the lowest (23). However, the dataset of this paper has not been published, and no special validation method was mentioned beyond the explanation provided.

In 2021, [18] proposed a method for malware attribution in the Internet of Things based on the TF-IDF method. The paper introduced the SMOTE-RF method for classification. The dataset comprised 2,389 malware samples from 7 APT groups: Lazarus, APT28, Operation C-Major, APT29, Dropping Elephant, Sandworm, and Naikon. Lazarus had the highest number of malware samples (1,060), while Naikon had the lowest (127). However, the dataset from this paper has not been publicly released, and no additional information regarding its validation method has been provided.

In 2021, [26] conducted a study on software attribution methods, focusing specifically on the datasets in this field. This paper is the only one that has provided a proper evaluation of the datasets available up to that time. This paper primarily focuses on reviewing research related to malware attribution and examines various aspects that have not been addressed with similar comprehensiveness in any other work. Of course, this research also has significant limitations, which we do not intend to address in this paper. In particular, one section of this paper reviews previous datasets. Notably, one of the paper's weaknesses is that, despite its title focusing on malware attribution, it also examines numerous datasets, including source code and non-malicious binaries, which are only tangentially related to malware attribution. However, the review of datasets specifically related to malware attribution in this paper is conducted with good accuracy, filling a gap in such evaluations until then. We now proceed to discuss the dataset presented in this paper as its contribution. This dataset, named APTClass, contains 15,660 malware samples. It is considered the richest dataset in this research area to date. The research systematically and methodically prepared this dataset, leveraging knowledge of the strengths and weaknesses of previous datasets. The steps taken include creating a list of APT groups, collecting information from open-source sources, extracting hashes and labels, removing duplicates, and validating samples using VirusTotal. The researchers standardized the names of APT groups extracted from six different sources to eliminate duplicates, resulting in 1,532 unique names.

This paper presents a novel approach to collecting APT group labels based on malware hashes. Initially, the text of PDF reports, YARA and IoC rules are extracted, and malware hashes are searched using regular expressions. Irrelevant words such as punctuation, stop words and hashes are removed from the text, and patterns like APT<number>, APT-C-<number>, ATK<number>, SIG<number>, and FIN<number> are searched for. The paper mentions that this method is employed only when the malware developer group can be determined from the report text; otherwise, N-gram Metadata Search and Keyword Search are utilized. The dataset comprises 15,660 labeled malware samples classified into 164 groups. Due to the large number of identified groups, group names are not mentioned here. However, the dataset includes various APT groups, with the highest number of malware belonging to the Sig17 group (4992) and the lowest to the Higaisa group (53). Additionally, the paper classifies the countries of APT groups and determines the number of malware in each country.

As of now, the dataset from this paper has not been publicly released. However, according to correspondence with the authors, the dataset will be made publicly available after the official publication of the paper.

In 2022, [21] presented a cybersecurity platform named CSKG4APT, based on a knowledge graph of APT groups information. The model extracts related entities of threat information by analyzing descriptive information and logical relationships contained in reports. The Diamond model and threat hunting are employed to provide an APT group tracking scheme. The dataset of this paper contains 1,041 technical reports related to 25 APT groups. Relying on information from technical reports serves as a validation method in itself. A key aspect of this study has been the comprehensive investigation into the geographical origins of APT groups, their targeted regions, the industries impacted, and the underlying motivations of the attacks.

In the same year, [27] presented a study considering different aspects of malware for attribution. The evaluation was conducted using five machine learning algorithms: Decision Tree, SVM, K-Nearest Neighbors, Multilayered Perceptron, and Fair Clustering. The dataset comprises 3,594 malware samples belonging to 12 APT groups. This dataset is sourced from [24] that was previously introduced and no additional steps were taken by this paper to prepare it.

Among the latest research in the field of attribution, [28] in 2023 introduced a mechanism named APTer, capable of predicting the next stages of an attack from advanced threat groups and attributing it to a specific group. This paper utilizes Decision Tree, SVM, and neural network learning methods. The dataset consists of 634,476 security threat alarms and 31 traffic features for training, and 13,785 threat alerts and 34 EDR-related features for testing. A summary of the datasets presented so far and their specifications is provided in Table 1.

The examination of available datasets highlights critical hurdles in crafting robust and pertinent datasets for malware attribution. To propel this field forward, researchers must confront these challenges head-on:

- **Standardizing APT Group Names**: Disparate sources frequently employ varying identifiers for identical APT groups. To avert misclassification, dataset curators must unify these labels, employing a consistent naming convention across all datasets.
- **Navigating Multi-Group Utilization**: Certain malware may serve multiple APT groups or emerge from collaborative efforts. It is imperative to discern the primary

TABLE 1. Comparison of existing datasets.

Paper	Publication year	Published publicly	ML/DL techniques	Num of samples	Num of APTs	Validation method
Rosenberg et al [3]	2017	×	Deep Neural Network	1,000	–	–
Hong et al [23]	2018	×	Decision Tree, Tree Bagging, Random Forest, Naive Bayes, Support Vector Machine	1,088	7	–
APTMalware [24]	2019	✓	–	3,594	5	–
Noor et al [19]	2019	×	Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Deep Neural Network	327 reports	–	Selection of technical reports from publicly published reports can be considered as a validation method for this dataset.
Meng et al [20]	2019	×	Natural Language Processing (NLP)	249 reports	12	Validation of reports is done manually by experts.
Haddadpajouh et al [22]	2020	×	Decision Tree	1,200	5	According to the extraction of the Cuckoo report from the samples, the correctness of the classification will be validated.
Laurenza et al [25]	2020	×	Isolation Forest	2,086	15	–
Li et al [18]	2021	×	SMOTE-RF	2,389	7	–
Gray et al [26]	2021	×	–	15,660	164	This research has used VirusTotal to validate aggregated samples.
Ren et al [21]	2022	×	–	1,041	25	Relying on the information of technical reports is considered a validation method in itself.
Sahoo et al [27]	2022	✓	Decision Tree, Support Vector Machine, K-Nearest Neighbors, Multilayer Perceptron, Fair Clustering	Dataset of [24]	–	–
Sachidananda et al [28]	2023	×	Decision Tree, Support Vector Machine, Neural Network	634,476 security threat alarms and 31 traffic features	–	–

developer, ensuring accurate model training. Moreover, clear delineation is necessary regarding whether the dataset encompasses malware developed by multiple groups.

- **Distinguishing Detection vs. Attribution Datasets**: It is vital to draw a distinction between datasets tailored for malware detection and those geared towards attribution. Attribution-focused datasets should encompass a broad spectrum of malware families for each APT group, contrasting with detection datasets, which may rely on samples from a single family. However, sourcing diverse samples, particularly for groups with limited malware samples, poses a notable challenge.
- **Ensuring Geographic Diversity**: Existing datasets often exhibit a bias towards APT groups from specific regions, potentially overlooking others engaged in multiple malware campaigns. Future endeavors should strive for geographic inclusivity, fostering a more comprehensive dataset landscape.

By tackling these obstacles head-on, researchers can forge more resilient datasets, fostering deeper insights and catalyzing advancements in malware attribution research.

III. APTRACKER: UNVEILING THE PINNACLE OF ANALYTICAL DATASETS FOR MALWARE ATTRIBUTION

This study embarks on a meticulous and exhaustive review and assessment of existing datasets, delving into their respective strengths and weaknesses. Moreover, it identifies the challenges plaguing researchers in this domain and endeavors to proffer viable solutions. Building upon this groundwork, the subsequent endeavor involves the presentation of a dataset endowed with the virtues of its predecessors while mitigating their shortcomings.

The imperative of a rich dataset spanning a broad spectrum in terms of malware samples, families, and APT groups cannot be overstated. Such a repository not only fosters avenues for future research but also enhances the precision and efficacy of malware attribution systems.

Among previous works, there are valuable datasets with important strengths. However, as explained in Section II, these datasets have serious limitations and weaknesses that we have attempted to address. The following limitations and weaknesses are notable:

- **Failure to Provide an Accurate Validation Method**: Accurate validation methods are essential for establishing the validity and reliability of dataset samples.

Unfortunately, some previous papers did not specify the validation methods they employed. Additionally, samples collected in some other datasets can be considered valid implicitly based on the credibility of the resources they use. Additionally, some papers only referenced simple, single-step methods, which, while somewhat reliable, are not comprehensive. Among the previous works, only [26] presented a suitable and robust validation method.

- **Non-disclosure of the Dataset**: One crucial factor in evaluating a dataset is its potential to serve as a basis for further research. Additionally, for comparing the accuracy of two methods, it is preferable that both use the same dataset. Future researches may seek to compare their results with previous ones based on the datasets used in those papers. However, many datasets are unfortunately not publicly available.
- **Few Samples Available in Some Datasets**: The limited number of malware samples in some datasets affects the accuracy of training and testing learning models based on these datasets. Additionally, some of these datasets contain samples from a relatively large number of APT groups, resulting in very few samples per group and exacerbating the issue. It should be noted that while the number of samples in some datasets may be sufficient for certain learning methods, when a dataset is intended to be used as a basis for malware attribution research, it must be applicable to various learning models. This consideration becomes even more critical for new learning methods that rely on large amounts of data.
- **Lack of Access to Malware Samples**: In some previous datasets, the actual malware samples are not available; instead, only related technical reports or security alerts have been aggregated.

In addition to the aforementioned weaknesses, there are also significant limitations discussed at the end of Section IV. In this dataset, we have attempted to address these weaknesses and limitations as thoroughly as possible.

Addressing this pressing need, this section delves into the description of the dataset proffered in this study. In essence, the dataset encompasses 22 APT groups, comprising 6,440 binary files of Windows malware samples from 40 distinct malware families. The APTs covered in this dataset, belong to 8 countries: Russia, South Korea, China, USA, Nigeria, North Korea, Pakistan and Belarus (refer to Table 2). It is worth noting that in certain instances, where data was sourced from previous datasets, the names of malware families associated with some APT groups were undisclosed. The next important point is that, for standardization, among the different names mentioned in various sources for an APT group, this paper provides one name as the standard. The reason for our choice among the different names was to select those that are more widely recognized and popular among cybersecurity experts and researchers. Of course, it is not necessary for other research groups to make the same choices. Additionally, other pseudonyms are mentioned to assist researchers as effectively as possible. Figure 1 shows a graph representation of APTracker.

As observed in preceding datasets, a primary challenge lies in validating malware samples—ascertaining their rightful classification within the malware family and, consequently, the APT group to which they belong. To confront this challenge, our methodology hinges on sourcing samples from reputable repositories endorsed by IT security researchers and threat analysts. Vx-Underground[3] [29] has curated a comprehensive array of malware samples from these trusted sources, meticulously categorized based on the year of their inclusion in the dataset. Among these sources, the compilation spanning from 2020 to 2024, encapsulated in [29], includes MalwareBazaar[4] [30].

Moreover, among the sources furnishing a meticulous classification of APT groups and malware samples attributed to them, MITRE ATT&CK [31] stands out as perhaps the most robust and reliable resource.

Our methodology for dataset collection commenced with the retrieval of all malware samples cataloged by [29] from MalwareBazaar. Subsequently, we cross-referenced the malware families associated with these samples against the roster of malware families affiliated with APT groups delineated in [31]. Samples were retained if they aligned with the designated malware families of APT groups, resulting in the aggregation of a substantial corpus of malware samples for each APT group. However, our efforts did not culminate there. In addition to MalwareBazaar, we leveraged other esteemed and credible repositories, notably InTheWild Collection[5] [32], to augment the dataset procured from the preceding step. While other reputable sources were perused, they contributed minimally to dataset enrichment.

Furthermore, in conjunction with the APT group listings furnished by MITRE ATT&CK, sources such as APT Groups and Operations[6] [33] and Malpedia[7] [34] were scrutinized to serve as a corroborative validation for [31], supplementing the APT group lists and facilitating the identification of the respective countries associated with APT groups. Additionally, previously published datasets like [24], renowned for their reliability in the realm of malware attribution, were also incorporated into our dataset.

Despite the provenance of malware samples from reputable sources, we undertook supplementary validation measures. Each malware sample underwent analysis within a Cuckoo environment, where they were subjected to scrutiny by an array of prominent anti-malware solutions. The resultant reports generated by this renowned sandbox served as the conclusive validation step for our dataset. The use of the Cuckoo sandbox, due to the excellent and wide features it

[3]https://vx-underground.org
[4]https://bazaar.abuse.ch
[5]https://vxunderground.org/Samples/InTheWild%20Collection
[6]https://apt.threattracking.com
[7]mhttps://alpedia.caad.fkie.fraunhofer.de

TABLE 2. Our proposed dataset, APTracker.

APT Group	AKA	Origin	Num of samples	Families
APT1	Brown Fox, Byzantine Candor, COMMENT PANDA, Comment Crew, Comment Group, G0006, GIF89a, Group 3, PLA Unit 61398, ShadyRAT, TG-8223	China	405	–
APT10	ATK41, BRONZE RIVERSIDE, CVNX, Cloud Hopper, G0045, Granite Taurus, HOGFISH, Menupass Team, POTASSIUM, Red Apollo, STONE PANDAD, TA429, happyyongzi	China	234	o.a. PlugX
APT21	HAMMER PANDA, NetTraveler, TEMP.Zhenbao	China	106	TravNet
APT28	APT-C-20, ATK5, Blue Athena, FANCY BEAR, FROZENLAKE, Fighting Ursa, Forest Blizzard, G0007, Grey-Cloud, Grizzly Steppe, Group 74, Group-4127, IRON TWILIGHT, ITG05, Pawn Storm, SIG40, SNAKEMACKEREL, STRONTIUM, Sednit, Sofacy, Swallowtail, T-APT-12, TA422, TG-4127, Tsar Team, TsarTeam, UAC-0028	Russia	258	–
APT29	ATK7, Blue Kitsune, BlueBravo, COZY BEAR, Cloaked Ursa, G0016, Grizzly Steppe, Group 100, IRON HEMLOCK, ITG11, Midnight Blizzard, Minidionis, Nobelium, SeaDuke, TA421, The Dukes, UAC-0029, YTTRIUM	Russia	313	PinchDuke, GeminiDuke, CosmicDuke, MiniDuke, CozyDuke, OnionDuke, SeaDuke, HammerDuke, CloudDuke, WellMail
APT30	G0013, LotusBlossom, RADIUM, Raspberry Typhoon	China	164	–
Dark Hotel	APT-C-06, ATK52, DUBNIUM, Dark Hotel, Fallout Team, G0012, Karba, Luder, Nemim, Nemin, Pioneer, SIG25, Shadow Crane, T-APT-02, TUNGSTEN BRIDGE, Tapaoux, Zigzag Hail	South Korea	273	Dark Hotel
Energetic Bear	ALLANITE, ATK6, BERSERK BEAR, BROMINE, Blue Kraken, CASTLE, Crouching Yeti, DYMALLOY, Dragonfly, G0035, Ghost Blizzard, Group 24, Havex, IRON LIBERTY, ITG15, Koala Team, TG-4192	Russia	132	Havex
Equation Group	EQGRP, G0020, Tilded Team	USA	395	FannyWorm
FIN6	ATK88, Camouflage Tempest, G0037, GOLD FRANKLIN, ITG08, MageCart Group 6, SKELETON SPIDER, TAAL, White Giant	–	251	LockerGoga, Maze
Gold SouthField	Pinchy Spider, Gold Garden	Russia	2,183	Sodin
Gorgon Group	ATK92, G0078, Gorgon Group, Pasty Gemini, Subaat	Pakistan	961	–
Indrik Spider	DEV-0243, EvilCorp, Manatee Tempest, UNC2165	Russia	8,000	Dridex
Lazarus	APT 38, APT-C-26, APT38, ATK117, ATK3, Andariel, Appleworm, BeagleBoyz, Bluenoroff, Bureau 121, COPERNICIUM, COVELLITE, Citrine Sleet, DEV-0139, DEV-1222, Dark Seoul, Diamond Sleet, G0032, G0082, Genie Spider, Group 77, Hastati Group, Hidden Cobra, Labyrinth Chollima, Lazarus, Lazarus group, NICKEL GLADSTONE, NewRomanic Cyber Army Team, Nickel Academy, Operation AppleJeus, Operation DarkSeoul, Operation GhostSecret, Operation Troy, Sapphire Sleet, Stardust Chollima, Subgroup: Bluenoroff, TA404, Unit 121, Whois Hacking Team, ZINC, Zinc	North Korea	1,365	Hoplight, HotCroissant, Manuscrypt, WannaCry
MoustachedBouncer	–	Belarus	280	Disco
RTM	G0048	Russia	85	–
Sandworm	APT44, Blue Echidna, ELECTRUM, FROZENBARENTS, G0034, IRIDIUM, IRON VIKING, Quedagh, Seashell Blizzard, TEMP.Noble, TeleBots, UAC-0082, UAC-0113, VOODOO BEAR	Russia	8,700	GoldenEye
SilverTerrior	–	Nigerea	14,700	AgentTesla, Lokibot, Nonocore
TA505	ATK103, CHIMBORAZO, DEV-0950, Dudear, FIN11, G0092, GOLD TAHOE, GRACEFUL SPIDER, Hive0065, Lace Tempest, SectorJ04, SectorJ04 Group, Spandex Tempest	Russia	300	Amadey, Azorult, FlawedGrace, Get2
TA551	ATK236, G0127, Monster Libra, Shakthak, GOLD CABIN	Russia	10,255	IcedlD, Gozi, Qbot
Winnti	–	China	390	Winnti
Wizard Spider	DEV-0193, DEV-0237, FIN12, GOLD BLACKBURN, Periwinkle Tempest, Pistachio Tempest, Storm-0193, TEMP.MixMaster, Trickbot LLC, UNC2053	Russia	14,690	Dyre, Emotet, Ryuk, TrickBot

offers, which were mentioned earlier, can be considered a solid and multifaceted validation method. Figure 2 delineates the comprehensive process entailed in dataset preparation.

It is worth noting that certain malware may be attributed to multiple APT groups as developers. In this research, we endeavored to discern and present the most probable developer by scrutinizing multiple sources and taking into account other malware attributed to APT groups.

In curating this dataset, we aimed to minimize inclusion of APT groups with either a paucity of associated malware samples or a lack of diversity in malware families, unless such circumstances accurately reflected real-world scenarios where no more samples were available. In instances where a substantial number of samples were attributed to an APT group or where the group was well-known, it remained included in the dataset, even if the variety of malware families within that group was limited. Anyway, the main reason some APT groups are absent from our dataset is the lack of sufficient malware samples from those groups.

As observed in previous datasets, one of the challenges in preparing a malware attribution dataset lies in the imbalance between the of malware samples across different APT groups, as well as within different malware families of the same group. There isn't a one-size-fits-all solution for addressing this issue. While employing a machine learning model may be effective for some datasets, it might not yield optimal results for others. It is plausible that different machine learning models need to be applied for various APT groups and malware families within those groups. Additionally, the choice of attribution features plays a crucial role and needs to be assessed independently for each research endeavor. However, delving into the specifics of this matter exceeds the scope of this research.

What's imperative in dataset preparation is that before automatic attribution using machine learning models, the aggregated samples should be examined through the lens of a malware analyst. To simulate the mindset of a malware attribution system, consider employing deep learning techniques, as a suitable method to extract pertinent features for attributing malware to APT groups. Even then, researchers must guide the deep learning model to some extent, directing it towards features believed to be influential. This is what's encapsulated by the term ''analytical data'' in the title of this paper.

Indeed, the necessity of a thorough analytical examination of the dataset becomes more apparent when there is an imbalance in the number of samples across different malware families within an aggregated set for an APT group. Consider an APT group whose malware samples are aggregated from three malware families. Let's suppose the first family comprises several thousand samples, the second family has a few hundred samples, and the third family has only a few tens of samples. Does such a distribution of malware samples imply overlooking the influence of the latter two families in establishing the APT pattern? Should we strive to balance the number of samples across all families, or should we consider employing separate machine learning models for each family, effectively creating a multi-stage malware attribution system? Or perhaps, despite the disparity in the number of samples, these three malware families share common characteristics that, when considered together, can reveal a pattern indicative of the APT group?

These are the questions that necessitate multiple iterations of testing in the realm of malware attribution research. There isn't a one-size-fits-all answer applicable to all models and features. What this paper aims to address is the introduction of features specific to malware families in the prepared dataset that, from the perspective of a malware analyst, can significantly impact the attribution of malware to APT groups. We have meticulously compiled these features for all malware families in the APTracker dataset, categorizing them coherently and making them available on our research GitHub page[8] [35]. Paying heed to these features can pave the way for future research endeavors.

The mentioned features include:

- **Tactics, Techniques, and Procedures (TTPs)**: This includes specific tactics, techniques, and procedures employed by each malware family. However, it is important to note that only those relevant to the malware itself are provided, excluding broader tactics such as early access related to the overall structure of the attack.
- **Type of Malware**: This feature denotes the category or type of malware within each family. It could include classifications such as trojan, RAT, spyware, ransomware, etc.
- **Programming Language**: Identifies the programming language(s) utilized in the development of the malware family. This information can offer insights into the skillset and preferences of the threat actors behind the APT group.
- **Comparison of TTPs**: This feature provides a comparative analysis of the tactics, techniques, and procedures observed across the malware families associated with each APT group. It helps in identifying commonalities and distinctions between different families within the same APT group, facilitating a deeper understanding of their operational methods.

These features collectively contribute to a comprehensive understanding of the malware landscape associated with various APT groups, enabling more accurate attribution and proactive threat mitigation strategies.

IV. REMAINING CHALLENGES

Previously, section II-B discussed the main challenges associated with preparing datasets in the field of attribution. This research aimed to address these challenges as effectively as possible during the dataset preparation process, ultimately providing a dataset that shows significant improvement compared to previous works.

[8]https://github.com/me-mazaheri/APTracker

FIGURE 1. APTracker dataset graph; APTs, country of origin, malware families, and malware types.

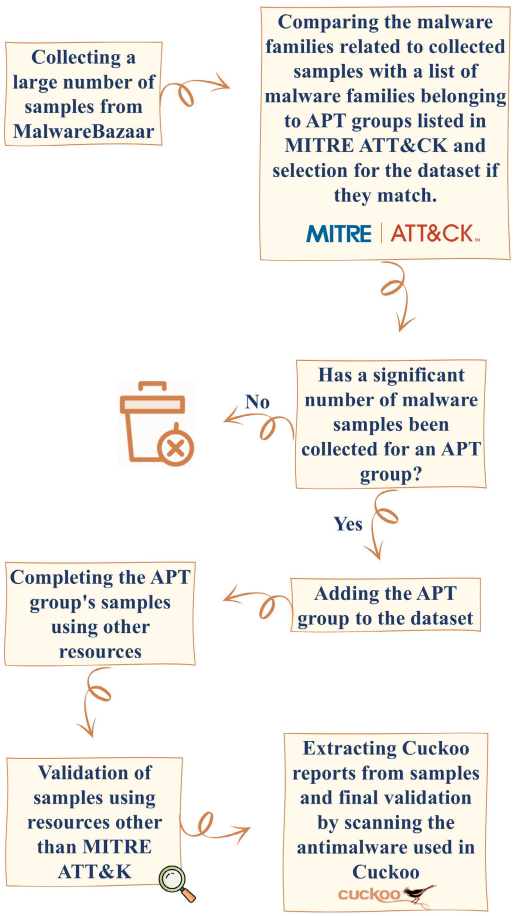

FIGURE 2. APTracker dataset preparation workflow.

However, it should be noted that, although a richer and more comprehensive dataset could still be developed, there are ongoing issues and challenges that could be explored in future research:

- **Attribution of malware developed by multiple groups**: As mentioned earlier, some malware has been developed by multiple APT groups. It is important to clarify that we are referring to malware that has been created by several groups from the outset, rather than malware used by multiple groups. This topic is not the focus of our paper. However, it appears that a valuable area for future research could be the attribution of such malware, alongside other malware developed by a single group.
- **Lack of coverage for some countries in the datasets**: This challenge was discussed earlier. Generally, datasets in the field of malware attribution contain samples from specific countries. Several factors contribute to this situation. One possible reason is that these datasets may be somewhat biased, which seems plausible. We have made efforts to address this challenge to some extent in our work, though there is still room for improvement. Another reason is that many attacks from certain countries are designed and executed in a targeted manner, focusing on specific targets and industries within specific countries. As a result, it is challenging to obtain a large number of malware samples from these countries, unlike attacks that are more indiscriminately designed and target a wide range of global victims, such as some well-known ransomware attacks. Additionally, many countries are less involved in cyberattacks, which limits the availability of malware samples from those regions. Nevertheless, there is still potential for improving the geographical balance of datasets.
- **Imbalance in datasets**: In this research, we placed significant emphasis on providing a dataset with an acceptable level of balance. As a result, we had to exclude some malware families with a limited number of samples. However, it is important to acknowledge that, regardless of the accuracy of the research, certain malware families may inherently have a limited number of available samples. Consequently, attribution methods based on learning techniques must employ solutions such as data duplication or similar approaches. Nevertheless, it remains appropriate for future work to strive for even greater balance in their datasets.

V. CONCLUSION

The primary focus of this paper is to conduct a comprehensive review of datasets utilized in malware attribution systems, along with a thorough evaluation. Subsequently, we address several challenges and limitations encountered in dataset collection within this field, introducing our proposed dataset, APTracker, along with its specifications, preparation method, and validation process. As previously mentioned, APTracker stands out as the most extensive dataset in the field of malware attribution, comprising 64,440 malware samples spanning 40 families associated with 22 APT groups. For researchers intending to delve into the realm of malware attribution and utilize this dataset in their future work, careful consideration of the features and machine learning methods employed for attribution is imperative when selecting malware samples from APTracker.

Despite the substantial investment of time and effort in providing the APTracker dataset and thorough review of numerous sources, it is evident that the potential for innovation in research pathways persists. The field of malware attribution is no exception to this rule, and it is our hope that forthcoming studies will strive to enrich datasets in this domain as much as possible. Additionally, we discuss the challenges and limitations inherent in preparing a suitable dataset for malware attribution, with each challenge presenting an opportunity for innovation in future datasets.

It is worth noting that, aside from the discussion on datasets, the field of malware attribution remains ripe for new research endeavors, which extend beyond the scope of this paper. This responsibility falls upon studies dedicated to comprehensively reviewing and examining all facets of malware attribution systems.

REFERENCES

[1] K. Yasar. *What is Malware? Prevention, Detection and How Attacks Work*. Accessed: May 2024. [Online]. Available: https://www.techtarget.com/searchsecurity/definition/malware
[2] *What is a Threat Actor? Types & Examples*. Accessed: May 2024. [Online]. Available: https://www.sentinelone.com/cybersecurity-101/threat-actor/
[3] I. Rosenberg, G. Sicard, and E. David, "DeepAPT: Nation-state APT attribution using end-to-end deep neural networks," in *Proc. Artif. Neural Netw. Mach. Learn. (ICANN)*, Alghero, Italy. Cham, Switzerland: Springer, Sep. 2017, pp. 91–99.
[4] Z. Wang, Z. Feng, and Z. Tian, "Neural representation learning based binary code authorship attribution," in *Proc. 11th EAI Int. Conf. Digit. Forensics Cyber Crime (ICDF)*, Boston, MA, USA. Cham, Switzerland: Springer, Oct. 2020, pp. 244–249.
[5] E. Quiring, A. Maier, and K. Rieck, "Misleading authorship attribution of source code using adversarial learning," in *Proc. 28th USENIX Secur. Symp. (USENIX Security)*, 2019, pp. 479–496.
[6] B. Alsulami, E. Dauber, R. Harang, S. Mancoridis, and R. Greenstadt, "Source code authorship attribution using long short-term memory based networks," in *Proc. 22nd Eur. Symp. Res. Comput. Secur. (ESORICS)*, Oslo, Norway. Cham, Switzerland: Springer, Sep. 2017, pp. 65–82.
[7] V. Kalgutkar, R. Kaur, H. Gonzalez, N. Stakhanova, and A. Matyukhina, "Code authorship attribution: Methods and challenges," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–36, Jan. 2020.
[8] X. Meng, B. P. Miller, and K. S. Jun, "Identifying multiple authors in a binary program," in *Proc. 22nd Eur. Symp. Res. Comput. Secur. (ESORICS)*, Oslo, Norway. Cham, Switzerland: Springer, Sep. 2017, pp. 286–304.
[9] Q. Liu, S. Ji, C. Liu, and C. Wu, "A practical black-box attack on source code authorship identification classifiers," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3620–3633, 2021.
[10] M. Abuhamad, T. AbuHmed, A. Mohaisen, and D. Nyang, "Large-scale and language-oblivious code authorship identification," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 101–114.
[11] F. Ullah, J. Wang, S. Jabbar, F. Al-Turjman, and M. Alazab, "Source code authorship attribution using hybrid approach of program dependence graph and deep learning model," *IEEE Access*, vol. 7, pp. 141987–141999, 2019.
[12] E. Bogomolov, V. Kovalenko, Y. Rebryk, A. Bacchelli, and T. Bryksin, "Authorship attribution of source code: A language-agnostic approach and applicability in software engineering," in *Proc. 29th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, Aug. 2021, pp. 932–944.
[13] S. Alrabaee, E. B. Karbab, L. Wang, and M. Debbabi, "BinEye: Towards efficient binary authorship characterization using deep learning," in *Proc. 24th Eur. Symp. Res. Comput. Secur. (ESORICS)*. Cham, Switzerland: Springer, Sep. 2327, pp. 47–67.
[14] A. Caliskan, F. Yamaguchi, E. Dauber, R. Harang, K. Rieck, R. Greenstadt, and A. Narayanan, "When coding style survives compilation: De-anonymizing programmers from executable binaries," 2015, *arXiv:1512.08546*.
[15] S. Alrabaee, N. Saleem, S. Preda, L. Wang, and M. Debbabi, "OBA2: An onion approach to binary code authorship attribution," *Digit. Invest.*, vol. 11, pp. S94–S103, May 2014.
[16] X. Meng, B. P. Miller, and S. Jha, "Adversarial binaries for authorship identification," 2018, *arXiv:1809.08316*.
[17] N. Rosenblum, X. Zhu, and B. P. Miller, "Who wrote this code? Identifying the authors of program binaries," in *Proc. 16th Eur. Symp. Res. Comput. Secur. (ESORICS)*, Leuven, Belgium. Berlin, Germany: Springer, Sep. 2011, pp. 172–189.
[18] S. Li, Q. Zhang, X. Wu, W. Han, and Z. Tian, "Attribution classification method of APT malware in IoT using machine learning techniques," *Secur. Commun. Netw.*, vol. 2021, pp. 1–12, Sep. 2021.
[19] U. Noor, Z. Anwar, T. Amjad, and K.-K.-R. Choo, "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise," *Future Gener. Comput. Syst.*, vol. 96, pp. 227–242, Jul. 2019.
[20] L. Perry, B. Shapira, and R. Puzis, "NO-DOUBT: Attack attribution based on threat intelligence reports," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 80–85.
[21] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, and Z. Tian, "CSKG4APT: A cybersecurity knowledge graph for advanced persistent threat organization attribution," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5695–5709, Jun. 2023.
[22] H. Haddadpajouh, A. Azmoodeh, A. Dehghantanha, and R. M. Parizi, "MVFCC: A multi-view fuzzy consensus clustering model for malware threat attribution," *IEEE Access*, vol. 8, pp. 139188–139198, 2020.
[23] J. Hong, S. Park, S. Kim, D. Kim, and W. Kim, "Classifying malwares for identification of author groups," *Concurrency Comput., Pract. Exper.*, vol. 30, no. 3, p. e4197, Feb. 2018.
[24] Cyber-Research. (2019). *APTMalware*. Accessed: May 2024. [Online]. Available: https://github.com/cyber-research/APTMalware
[25] G. Laurenza, R. Lazzeretti, and L. Mazzotti, "Malware triage for early identification of advanced persistent threat activities," *Digit. Threats, Res. Pract.*, vol. 1, no. 3, pp. 1–17, Sep. 2020.
[26] J. Gray, D. Sgandurra, and L. Cavallaro, "Identifying authorship style in malicious binaries: Techniques, challenges & datasets," 2021, *arXiv:2101.06124*.
[27] D. Sahoo, "Cyber threat attribution with multi-view heuristic analysis," in *Handbook of Big Data Analytics and Forensics*. Cham, Switzerland: Springer, 2022, pp. 53–73.
[28] V. Sachidananda, R. Patil, A. Sachdeva, K. Y. Lam, and L. Yang, "APTer: Towards the investigation of APT attribution," in *Proc. IEEE Conf. Dependable Secure Comput. (DSC)*, Nov. 2023, pp. 1–10.
[29] *VX-Underground*. Accessed: May 2024. [Online]. Available: https://www.vx-underground.org
[30] *MalwareBazzar*. Accessed: May 2024. [Online]. Available: https://www.bazaar.abuse.ch
[31] *MITRE ATT&CK*. Accessed: May 2024. [Online]. Available: https://www.attack.mitre.org
[32] *In The Wild Collection*. Accessed: May 2024. [Online]. Available: https://vx-underground.org/Samples/InTheWild
[33] *APT Groups and Operations*. Accessed: May 2024. [Online]. Available: https://apt.threattracking.com
[34] *Malpedia*. Accessed: May 2024. [Online]. Available: https://malpedia.caad.fkie.fraunhofer.de/
[35] M. E. Mazaheri. (2024). *APTracker*. Accessed: Aug. 2024. [Online]. Available: https://github.com/me-mazaheri/APTracker

MOHAMAD ERFAN MAZAHERI received the B.Sc. degree from Bu-Ali Sina University and the M.Sc. degree in computer engineering from the Sharif University of Technology. He is currently pursuing the Ph.D. degree in computer engineering with Shahid Beheshti University. His research interests include information security, malware detection and attribution, threat hunting, and microarchitectural attacks.

ALIREZA SHAMELI-SENDI received the B.Sc. and M.Sc. degrees from the Amirkabir University of Technology and the Ph.D. degree in computer engineering from Montreal University (Ecole Polytechnique de Montreal), Canada. He is currently an Associate Professor with Shahid Beheshti University (SBU). Before joining SBU, he was a Postdoctoral Fellow with Ericsson, Canada, and a Postdoctoral Researcher with ETS and McGill universities in collaboration with Ericsson. His research interests include information security, intrusion response systems, and cloud computing. He was a recipient of the Postdoctoral Research Fellowship Award and the Industrial Postdoctoral Fellowship Award from Canada. In addition, he received the Best Researcher Awards at SBU, in 2018 and 2022.