Current Challenges of Cyber Threat and Vulnerability Identification Using Public Enumerations

Lukáš Sadlek
Masaryk University
Brno, Czech Republic
sadlek@mail.muni.cz

Pavel Čeleda
Masaryk University
Brno, Czech Republic
celeda@ics.muni.cz

Daniel Tovarňák
Masaryk University
Brno, Czech Republic
tovarnak@ics.muni.cz

ABSTRACT

Identification of cyber threats is one of the essential tasks for security teams. Currently, cyber threats can be identified using knowledge organized into various formats, enumerations, and knowledge bases. This paper studies the current challenges of identifying vulnerabilities and threats in cyberspace using enumerations and data about assets. Although enumerations are used in practice, we point out several issues that still decrease the quality of vulnerability and threat identification. Since vulnerability identification methods are based on network monitoring and agents, the issues are related to the asset discovery, the precision of vulnerability discovery, and the amount of data. On the other hand, threat identification utilizes graph-based, nature-language, machine-learning, and ontological approaches. The current trend is to propose methods that utilize tactics, techniques, and procedures instead of low-level indicators of compromise to make cyber threat identification more mature. Cooperation between standards from threat, vulnerability, and asset management is also an unresolved issue confirmed by analyzing relationships between public enumerations and knowledge bases. Last, we studied the usability of techniques from the MITRE ATT&CK knowledge base for threat modeling using network monitoring to capture data. Although network traffic is not the most used data source, it allows the modeling of almost all tactics from the MITRE ATT&CK.

CCS CONCEPTS

• **Security and privacy** → **Network security**; **Vulnerability management**; • **Networks** → *Network monitoring*.

KEYWORDS

cyber threat intelligence, threat identification, vulnerability identification, CVE, CAPEC, MITRE ATT&CK

1 INTRODUCTION

Threat identification is the task of revealing any possible event jeopardizing organization's assets. Several enumerations and knowledge bases support such a use case by defining vocabulary for specific areas of cybersecurity [12]. *Common Vulnerabilities and Exposures* (CVE) [6] define how vulnerability identifiers are assigned to discovered vulnerabilities. *Common Weakness Enumeration* (CWE) [7] categorizes common weaknesses, while *Common Platform Enumeration* (CPE) [5] defines a way of assigning standardized identifiers to classes of technological assets. Last, *Common Attack Pattern Enumeration and Classification* (CAPEC) [8] classifies common attack patterns. A related knowledge base *MITRE ATT&CK* [9] provides a matrix of adversarial tactics and techniques.

These enumerations are used within the *cyber threat intelligence* (CTI) that provides knowledge about goals, methods, previous attacks, and current attack possibilities of adversaries in cyberspace. Even though more than half of the *SANS CTI survey* respondents use CTI for threat and vulnerability management [4], only 39% of organizations knew that they were vulnerable prior to the data breach based on Ponemon Institute's study [36]. These statistics show that current approaches for threat and vulnerability identification may not be perfect.

These enumerations belong to the threat, vulnerability, and asset management. Therefore, the interoperability of these managements should support the identification of cyber threats. For example, *NIST Cybersecurity Framework* recommends several activities supporting risk evaluation of cyber threats caused by vulnerable assets, optionally using shared CTI [31], such as public enumerations. Data from public enumerations need to be enriched with data about the organization's assets obtained from network monitoring, network scanning, and application and system logs.

In this paper, we focus on identifying threats and vulnerabilities in cyberspace using the aforementioned public enumerations and data about assets. For the simplicity of expression, we also consider the ATT&CK knowledge base when using term enumerations in this paper. We propose the following research questions:

(1) *What are the current challenges of vulnerability and cyber threat identification using enumerations and data about assets?*
(2) *What is the usability of MITRE ATT&CK for threat modeling when only network monitoring is used as a source of data?*
(3) *What is the interoperability of public enumerations using references between their entries?*

In the first research question, we summarize the current state of the art. We focus on methods that identify enumeration entries, excluding methods that discover only some entries using their related detection methods. In the second research question, we study whether the well-known MITRE ATT&CK knowledge base can

be used for cyber threat identification using network monitoring. Finally, the third research question evaluates the interoperability of the enumerations using their references.

The paper is divided into six sections. First, we describe related work in Section 2. Section 3 contains state of the art for vulnerability identification and challenges of the current state. Section 4 describes the identification of threats using public enumerations and knowledge bases, including the current challenges in this area. The following Section 5 describes the results of the analysis of public enumerations. Last, Section 6 concludes the paper.

2 RELATED WORK

The related work consists of several parts related to the threat and vulnerability identification. The first part covers enumerations actively used in this research area. The following part introduces the position of these enumerations within the field of the *cyber threat intelligence* (CTI). The last part justifies that related work did not cover this area of research.

2.1 Security Enumerations

The enumerations considered in this paper belong to different parts of the security operations management – asset, vulnerability, and threat management (see Figure 1). An enumeration belonging to asset management is called the *Common Platform Enumeration* (CPE) [5]. The well-known enumerations that belong to vulnerability management are the *Common Vulnerabilities and Exposures* (CVE), listing discovered vulnerabilities, and the *Common Weakness Enumeration* (CWE), listing commonly appearing weaknesses. *Common Attack Pattern Enumeration and Classification* (CAPEC) [8] and *MITRE ATT&CK* [9] belong to threat management.

The CPE is a standard for naming classes of products – hardware devices, operating systems, and applications. These classes with other product information are specified within a formatted string called the CPE match string used in data feeds from the *National Vulnerability Database (NVD)* [32].

The current version, CPE 2.3, uses syntax expressed in Figure 2 [5]. The most important parts are at the beginning of the string. The *part* section has three possible values: *a* for applications, *o* for operating systems, and *h* for hardware. The *vendor* expresses the name of the product's vendor. The product name is filled in the *product* section and followed by the product's version. The CPE match string may also contain the product's update, edition, and other parts [5]. The syntax allows specifying special characters, e.g., an asterisk is used to denote *ANY* value in the appropriate part of the formatted string.

Figure 1: Mapping of data entities about assets, vulnerabilities, and threats.

cpe:2.3:<part>:<vendor>:<product>:<version>:<update>:<edition>: :<language>:<sw_edition>:<target_sw>:<target_hw>:<other>

cpe:2.3:o:debian:debian_linux:11.0:*:*:*:*:*:*
cpe:2.3:a:apache:log4j:2.0:rc1:*:*:*:*:*:*

Figure 2: CPE match string and its example content.

The *Common Vulnerabilities and Exposures* (CVE) [6] is a list of vulnerabilities published in the *National Vulnerability Database* (NVD) [32] maintained by MITRE. Each CVE entry contains an identification number, description, and URL references to sources where the vulnerability was published, e.g., on vendor websites. The NVD stores CVE vulnerabilities also with their weaknesses (CWE ID) and vulnerable products (CPE).

An example of a CPE match string for CVE vulnerability is the *log4j's* match string associated with the *Log4Shell* vulnerability (see Figure 2). The NVD [32] also specifies CPE configurations consisting of several CPEs grouped by keywords *AND* and *OR*. In such a way, the NVD specifies that CVE is present, e.g., on a specific operating system *AND* application installed on it. Using the *OR* keyword specifies several vulnerable CPE configurations.

The *Common Weakness Enumeration* (CWE) [7] lists common weaknesses in software or hardware products. This enumeration is organized as a tree with well-known examples of weaknesses, such as *CWE-81 (XML Injection)* and *CWE-94 (Code Injection)*. The entries also contain additional information, such as descriptions and mitigation methods. Related CWEs are assigned to each vulnerability in the NVD, e.g., *Log4Shell* vulnerability (CVE-2021-44228) is categorized as *CWE-502* (Deserialization of Untrusted Data). CWE also contains a mapping from *OWASP Top Ten 2021*.

CAPEC describes attack patterns and ATT&CK adversarial tactics and techniques. CAPEC is organized as a hierarchical tree of categories, e.g., *CAPEC-125 (Flooding)* is a parent of *CAPEC-482 (TCP Flood)*, *CAPEC-486 (UDP Flood)*, and other subcategories of the flood. On the contrary, MITRE ATT&CK is organized as a matrix where columns are represented by tactics and rows by techniques. Examples of ATT&CK techniques are *Phishing (T1566)* and *Brute Force (T1110)*. Examples of references between data entries from Figure 1 are *CAPEC-98* and the ATT&CK technique with ID *T1566*. Both of them reference each other and express phishing.

The *Structured Threat Information Expression* (STIX) is a format for expressing and serialization of CTI [35]. STIX represents a threat as a graph with CTI entities (e.g., attack pattern, vulnerability, malware, and course of action) and their relationships. The STIX allows storing entries from the enumerations in its objects. For example, *attack pattern* objects may contain CAPEC and ATT&CK entries, while *vulnerability* objects often contain CVE vulnerabilities.

Sauerwein et al. [43] conducted a triangulation study about public security data sources. According to the authors, information about vulnerabilities and attacks or vulnerabilities and threats often appears in one data source. They found out that CVE and STIX are two primary standards that are used. CVE identifiers appear almost in every second data source, CWE in 13% of them, while the other standardized enumerators rarely appear [43]. Therefore, the task of cyber threat and vulnerability assessment using the
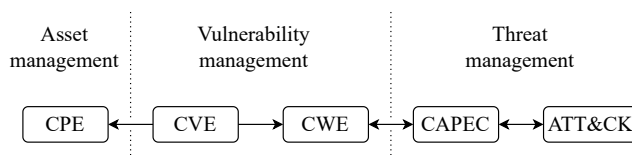
Table 1: Comparison of structure and sources of mentioned enumerations.

Source	Structure	Free-Text Fields	Data Source
CVE	flat (list)	partially	NVD
CPE	hierarchical (string)	no	NVD
CWE	hierarchical (tree)	partially	MITRE website
CAPEC	hierarchical (tree)	partially	MITRE website, Github
ATT&CK	hierarchical (matrix)	partially	Github

Table 2: Comparison of levels from DML model and PoP.

	Detection Maturity Level	Pyramid of Pain
1	atomic indicators	domain names IP addresses hash values
2	host & network artifacts	network & host artifacts
3	tools	tools
4	procedures	
5	techniques	TTPs
6	tactics	
7	strategy	
8	goals	——
9	identity	

enumerations is not trivial and has its rationale. A summary of mentioned enumerations is in Table 1.

2.2 Position of Enumerations in Cyber Threat Intelligence

Enumeration entries can be categorized into several CTI categories according to their maturity. CTI is generally related to indicators of compromise (IOCs) that can be divided into atomic, computed, and behavioral [20]. Atomic indicators cannot be divided anymore, e.g., IP addresses or CVE IDs. On the contrary, computed indicators are literally computed from the captured data, e.g., file hash values. Lastly, behavioral indicators describe the attacker's behavior using atomic and computed indicators [20]. E.g., ATT&CK and CAPEC IDs describing the attacker's activity are behavioral indicators.

Relationships among distinct parts of CTI can also be expressed using the *Pyramid of Pain (PoP)* [1] and the *Detection Maturity Level (DML)* model [45] or its extended version [26]. The particular PoP levels express the attacker's difficulties when the defense operates with more sophisticated indicator types. Similarly, the DML model expresses the organization's maturity in applying different indicator types, from the most low-level to the most abstract. The DML model and the PoP are similar, but the PoP divides atomic indicators into three different levels (see Table 2). Tactics, techniques, and procedures (TTPs) that can be obtained, e.g., from MITRE ATT&CK, express the most mature indicators used for the security defense.

2.3 Cyber Threat Intelligence Surveys

In this section, we justify that our research area is not described sufficiently by the related work since our paper's limitation is an absence of methodological literature search. In general, the existing surveys study the area of CTI without focusing on identifying cyber threats and vulnerabilities using the enumerations.

Tounsi and Rais [48] surveyed issues and emerging trends of technical threat intelligence. They pointed out that the traditional approaches in this area are not sufficient to cope with advanced persistent threats, zero-day threats, composite threats, or polymorphic threats. *Mavroeidis and Bromander* [26] evaluated CTI taxonomies, sharing standards, and ontologies. They introduced their CTI model

since there was no complete ontology covering all types of CTI data. *Wagner et al.* [50] surveyed the current state of CTI sharing. They claim that it is challenging to participate in CTI sharing because of insufficient models and tools.

Threat modeling develops threat representation and analyzes adversarial possibilities. *Bodeau et al.* [2] prepared a technical report about cyber threat modeling comparing characteristics of threat models and frameworks. *Xiong and Lagerström* [52] conducted a systematic literature review about threat modeling. They found that threat modeling is diverse, but most activities are manually accomplished. A promising research direction is to connect threat models with threat and vulnerability databases.

Other surveys may seem to be related to this paper but do not focus on identifying vulnerabilities and cyber threats to the organization's assets. *Le et al.* [24] surveyed data-driven vulnerability assessment using machine learning and deep learning. They focused on the prioritization and the prediction of exploitation, impact, severity, type, and other properties of vulnerabilities. *Rahman et al.* [39] provided a survey of threat intelligence extraction from text. The primary goal was to extract important entities, not to identify threats in the organization. Therefore, only some papers listed in the survey are relevant to our topic. *Kaloudi and Li* [22] surveyed the cyber threat landscape related to artificial intelligence but without focusing on the enumerations.

3 VULNERABILITY IDENTIFICATION

Vulnerability identification using public enumerations is a task to reveal the presence of a published vulnerability in the organization. The task can be accomplished in two manners. The first one is based on asset discovery and vulnerability databases. The second one uses published exploits to reveal vulnerable systems. It is more reliable and is often used as *vulnerability validation* after applying the first one.

3.1 State of the Art

Vulnerability identification methods based on asset discovery may obtain information about assets from passive network monitoring, active network monitoring, and agents. Passive network monitoring consists of capturing network traffic at some observation point [18] without any modification of traffic. On the contrary, active network

monitoring targets artificial requests to network services using network scanners [3]. Agent-based approaches use software checking installed applications on devices and their versions. In the following text, we will focus on vulnerability identification approaches built upon these data collection methods.

The task of identifying CVE vulnerability using data about assets can be divided into two subtasks – construct the CPE match string and find the appropriate CVE from the NVD using the constructed string. The first subtask can be accomplished using banner grabbing and OS and network service fingerprinting if the network monitoring is used.

Banner grabbing is a technique of obtaining responses from services running on opened ports when a network connection is initiated. This response is parsed to extract parts of the CPE match string in Figure 2. Usually, such a method can provide part, vendor, product, version, and sometimes even other fields of the string.

O'Hare et al. [34] used metadata obtained from the Censys scanner that parses service banners to obtain the list of CPEs representing vulnerable services. *Genge and Enăchescu* [17] used the results of Shodan queries. They extracted vendor names, product names, and product versions from service banners to identify relevant vulnerabilities from the NVD. Shodan and vulnerabilities from the NVD were also used by *Samtani et al.* [42].

OS and network service fingerprinting are other options for constructing the CPE match string. OS fingerprinting captures network connection properties, such as TCP window size and Time to Live (TTL), to infer the device's operating system. Passive OS fingerprinting from IP flow was used by *Laštovička et al.* [23] to enumerate vulnerabilities in large networks.

Banner grabbing and fingerprinting approaches are often implemented in network scanners, e.g., Nmap [25] and Shodan. The Nmap scanner outputs the CPE string containing vendor name, product name, and product version of scanned network services.

The network monitoring can be substituted by system and application logs to identify products installed on devices. An approach similar to agent-based approaches was proposed by *Gawron et al.* They utilized Unix system logs, Windows log events, web server logs, and proxy logs to extract CPE identifiers of applications [16].

The second subtask is to determine CVE using the constructed string. Related work mentions several options for this subtask. Linear search through the CPE dictionary is the most straightforward option. For example, sequential lookup and Levenshtein distance were used by *O'Hare et al.* [34]. *Genge and Enăchescu* [17] introduced a more mature approach. They used the hash tables for storing possible CPE names and a tree for hierarchically storing version numbers. Bipartite graphs were then used to model the relationship between CVEs and related CPEs, while hyper edges modeled CPE configurations (see Section 2).

An approach that does not use data about assets but checks a published exploit or presence of weakness is implemented in vulnerability scanners. Some scanners (e.g., Arachni) can identify weaknesses such as XSS, SQL, and code injections, and output discovered weaknesses as CWE and CAPEC entries. Nessus identifies network services, operating systems, and vulnerabilities [46]. Metasploit is an example tool [40] that validates the presence of vulnerabilities using exploits from Exploit DB [44]. Some of these exploits even have their CVE ID assigned.

3.2 Challenges

Challenges of vulnerability identification using public enumerations arise from the relationship of vulnerability management with asset management, shortcomings of current methods, and a large amount of data.

Asset Management: Vulnerability identification is associated with situational awareness about vulnerable assets in the organization. However, according to Edgescan's 2020 Vulnerability statistics report, most professionals do not know about all organization's assets, e.g., web applications or endpoints [14]. When methods based on asset discovery are used, it may cause some organizations to successfully patch vulnerabilities on known assets while leaving the unknown completely unpatched [51].

Vulnerability Discovery Precision: Another issue of vulnerability discovery is a lot of false positives. Methods that have extensive coverage (e.g., based on OS fingerprinting) can be imprecise or will not reveal sufficient details [23]. It causes employees to spend much time validating vulnerabilities [14], e.g., using vulnerability scanners that can use automated vulnerability exploits to confirm true positives.

In our scope, this issue is caused by constructing a CPE string that contains many *ANY* values. *Gawron et al.* [16] can often extract only an application name and a version from logs for the CPE match strings (see Figure 2). On the contrary, *Laštovička et al.* [23] used active scanning and passive monitoring that can determine only the first parts of the match string up to the version part.

Evaluation can be found in several papers. Examples of discussed approaches are implemented in Scout, which has 73 % precision [34], and ShoVAT, which outputs 7.77 % false positives [17]. Researchers usually use the results of other vulnerability scanners or approaches as ground truth for their methods [17, 23, 34].

Amount of Data: The NVD currently contains more than 170 thousand vulnerabilities. More than 8000 entries were added to the CPE dictionary during each month of 2021 [33]. Currently, implemented approaches are still efficient enough to be used in practice. However, timely aspects and asset landscape visibility are becoming more critical with the increasing amount of published vulnerabilities.

Implementation of CPE Specifications: CPE has several specifications that standardize CPE naming, CPE dictionary, CPE applicability language, and CPE name matching [49]. However, only specification for CPE naming of products [5] is usually used in research papers. These research papers implement custom algorithms for CPE matching. A graph-based approach that conforms to all four specifications was provided in [49].

3.3 Possible Research Directions

We identified two possible research directions based on the current state of the art. Since commercial solutions for vulnerability management consider vulnerability identification one of its essential tasks, their creators are motivated to advance the current state. However, we still need approaches backed by scientific evaluations.

Interoperability: The first possible research direction is to create a functioning system with good network visibility by unifying

partial approaches. This research direction has its rationale since vulnerability identification approaches have some advantages. They were evaluated by *Laštovička et al.* [23] who showed that the passive network monitoring reveals more vulnerable hosts assuming a particular OS fingerprinting method can process encrypted traffic. However, active network monitoring has better accuracy because of sending specially crafted probes to identify selected services. Commercial solutions also unify the mentioned approaches (active monitoring, passive monitoring, and agents) to achieve higher accuracy of vulnerability identification for customers [38, 47].

Current IT Environments: Since current IT environments consist of on-premises and cloud assets, networks lose their traditional perimeters. Listed academic literature also does not focus on other novel aspects of IT environments, e.g., encrypted network traffic that hinders the use of some methods. However, commercial solutions for vulnerability management can cope with these environments and work with on-premises assets, cloud assets, and mobile endpoints [38, 47]. Their disadvantage is the missing detailed insight into these paid tools.

4 THREAT IDENTIFICATION

Cyber threat identification belongs to essential tasks of threat management. For example, Cisco assigns capabilities such as threat intelligence services to the fourth generation of security operations centers [29]. Cyber threat identification using the enumerations determines whether some categorized threat could materialize according to the current security posture. Naive identification of such threats correlates indicators of compromise with data obtained from the organization. However, the use of enumerations usually leads to more sophisticated methods that utilize *Tactics, Techniques, and Procedures (TTPs)*.

4.1 State of the Art

Enumeration entries can appear in CTI reports [19], within data models of cyber threats [27, 53], and be extracted from system and application logs [28]. Peer-reviewed literature provides several types of approaches that use the enumerations. Graph-based approaches can model entities (e.g., events) and their relationships [28]. Natural Language Processing (NLP) processes human-readable threat intelligence [19, 53] and can accomplish additional tasks with machine learning (ML), e.g., classification [30]. Last, ontological approaches help model the CTI domain [27, 37].

Graph-based approaches are helpful for modeling sequences of steps related to TTPs. These steps may be accomplished simultaneously with other legitimate activities of users, and attackers may skip or add some steps. Therefore, the rational method is to search for similarities with malicious TTPs. Such approaches may provide more granular results compared to simple detection mechanisms.

A graph-based approach for the detection of Advanced Persistent Threats (APTs) was proposed by *Milajerdi et al.* [28] based on a provenance graph from system audit logs and a high-level scenario graph containing APT's activities (TTPs) mapped to the cyber kill chain. The TTPs are found in the provenance graph and mapped to their kill chain phases. Each of the TTPs has its assigned severity from CAPEC. The final score is computed based on all kill chain phases to distinguish attacks with sufficient confidence.

The NLP is applied to automatically extract the data entities (e.g., TTPs, IOCs, domain tags) from human-readable CTI texts. However, this approach also requires a custom or a standardized taxonomy of entities that can be extracted, e.g., CAPEC and ATT&CK for TTPs [19]. This area contains a lot of scientific literature that describes extraction methods [39] but does not provide means to identify relevant threats to the protected assets. *Zhao et al.* [53] extracted CTI from social data using NLP and a convolutional neural network to identify CVE and other indicators from the CTI. Relevant threats are determined according to domain tags that describe where the organization belongs, e.g., government.

Machine-learning classification can be applied when extracted CTI entities can be ambiguous. *Noor et al.* [30] trained naive Bayes classifier processing incident descriptions to determine TTPs from ATT&CK. A semantic network from threats and the TTPs allows determining relevant threats according to the host and network artifacts belonging to the second level of DML in Section 2.2.

Ontological approaches are usually based on a representation of CTI using standards, taxonomies, threat reports, and network architecture. Populated ontologies allow inferring new relationships among their entities. The CTI ontology was used by *Mavroeidis and Jøsang* [27] for threat hunting from Sysmon logs. The ontology contains, e.g., threat actors, attack patterns, and indicators of compromise. It can reveal the attacker's goals and suitable countermeasures. *Qamar et al.* [37] created ontology that maps threats to the network based on STIX reports, CVE, and network architecture. The inference rules determine relevant threats and identify affected assets based on the CTI, its TTPs, indicators, and vulnerabilities.

4.2 Challenges

Challenges of threat identification using CTI arise from the relationship of threat management with asset and vulnerability managements, a large amount of data, unstructured data, and shortcomings of current methods.

Unstructured CTI Reports: Unstructured text of CTI reports causes difficulties with automated processing. One of the issues is to extract relevant entities [39], e.g., an attacker's name, techniques, and exploited vulnerabilities. Standardized enumerations and knowledge bases provide taxonomies of obtainable output, e.g., vulnerability ID and name of an attack technique. Therefore, researchers often utilize text mining and NLP [53].

Lack of Visibility and Amount of Data: Organizations usually obtain data from several external sources to identify cyber threats [4]. Therefore, the organizations cope with a large amount of CTI obtained from different sources and with different formats. It is also necessary to add internal context about assets. As a result, researchers aim to provide means of determining relevant threats [30, 37, 53].

Maturity of Methods: Naive methods correlate suspicious IOCs from CTI reports with data from the organization. However, such a method's usability is limited because the attacker can easily and quickly change these indicators, e.g., the validity of IP addresses quickly decreases after one day [48]. Therefore, researchers introduce methods that support TTPs [27, 28, 30, 37, 53] to elevate the methods to higher levels of DML. Researchers usually evaluate

methods on past CTI reports [37] or prepare some dataset containing threat-related activity, e.g., from a cybersecurity exercise [28].

4.3 Possible Research Directions

We identified two possible research directions related to this topic. The first one is the same as for the vulnerability identification, and the second one is related to machine learning. Both may be partially related to the current topic of SOAR (*Security Orchestration, Automation, and Response*) platforms [15].

Interoperability of Partial Solutions: The task is to research how to combine partial solutions to provide a functional system able to identify a wide range of cyber threats using the enumerations. It would be necessary to use several data sources [15] and methods that accomplish more than a direct combination of security alerts.

Machine Learning for Threat Identification: Machine learning provides several methods that can be applied for threat identification using the enumerations, such as classification and clustering. Some commercial tools apply machine learning within threat management, e.g., graph-based machine learning [11] and artificial intelligence [21]. However, the use of artificial intelligence is a threat itself [22].

5 ANALYSIS OF ENUMERATIONS

This section provides results that answer the second and the third research question. First, we parsed the ATT&CK matrix expressed in the STIX format from the MITRE's CTI repository [10]. This matrix contained tactics, techniques, data sources, countermeasures, adversary groups, malware, and their relationships. Each attack technique is related to data sources providing relevant information for its detection.

We studied how many ATT&CK techniques in the Enterprise ATT&CK matrix of version 10.1 are visible on the network level, i.e., they have been assigned as data source one of: *Network Traffic: Network Connection Creation*, *Network Traffic: Network Traffic Content*, and *Network Traffic: Network Traffic Flow*. We found that network traffic is relevant for 131 ATT&CK techniques and sub-techniques from the overall amount of 707 techniques and sub-techniques.

These techniques belong to 13 out of 14 phases of the ATT&CK kill chain, i.e., 13 tactics. Therefore, MITRE ATT&CK seems useful for modeling cyber threats when network monitoring is used as a data source. Such modeling may lead to methods operating on a sufficiently high level of maturity, as explained in Section 2.2.

The overall analysis of data sources in ATT&CK showed that the first place among all data sources belongs to *Command* (256 techniques and sub-techniques) and *Process* (253 techniques and sub-techniques), followed by *File* (192) and *Network Traffic* (131). There is a considerable gap between *Network Traffic* and the fifth data source – *Windows Registry* (see Table 3). Many techniques have no data source assigned.

We used references of enumerations denoted by arrows in Figure 1 to answer the third research question. Each entry may have zero or more references to other entries. Consistent use of these references can allow expressing the CTI, e.g., using enumeration entries represented by several STIX objects. For example, a newly discovered vulnerability (CVE) can be exploited by several attack

Table 3: The most common data sources in MITRE ATT&CK (22nd March 2022).

| Data Source	Count of Techniques
Command	256
Process	253
File	192
Network Traffic	131
Windows Registry	69
Application Log	55
Module	50

patterns (CAPEC, ATT&CK). It would be beneficial to use CAPEC or ATT&CK entries as detailed categories for vulnerability exploits.

Therefore, we took references of CVEs to their respective weaknesses (CWE). We complemented them with CWE entries' references to attack patterns (CAPEC). Finally, we added references of CAPEC entries to ATT&CK techniques. Figure 3 shows that approximately 13% of all vulnerabilities from the NVD do not have any related CAPEC. Then, 28.52% of vulnerabilities have some meaningful count of CAPECs (one to five related CAPECs), and the remaining vulnerabilities have a higher count of CAPECs. As expected, more than 73% of CVEs do not have any related ATT&CK technique because CWE is application-oriented, and ATT&CK is designated for network security (see Figure 4). However, approximately a quarter of CVEs have some ATT&CK technique, and more than 15% have a meaningful count of them.

These results reveal that we cannot automatically determine which attack pattern or attack technique will be related to the vulnerability exploitation using references between enumerations in most cases. In other words, the CVE vulnerabilities from the NVD prefer such a subset of CWEs that does not allow mapping to a reasonable amount of CAPEC and ATT&CK entries.

A more detailed analysis of the references also revealed that only approximately one fourth of CWEs references some CAPECs. When CWE entry is related to one ATT&CK technique, it usually has several CAPECs. It is mapped to only one ATT&CK technique because only one out of several CAPECs is mapped to that ATT&CK technique. Even if CWE transitively references one ATT&CK technique, it does not necessarily lead to this ATT&CK entry.

References between CAPEC and ATT&CK can also be analyzed in the opposite way. Analysis of network-related ATT&CK techniques at the beginning of this section revealed that only 48 out of
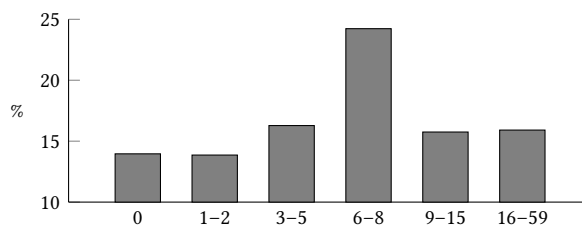
Figure 3: How many CAPEC entries mapped to one CVE (15th January 2022).
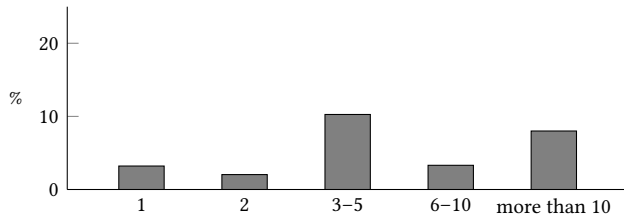
Figure 4: How many ATT&CK techniques mapped to one CVE (15th January 2022).

these 131 techniques and sub-techniques have at least one assigned CAPEC. These ATT&CK entries reference 22 CAPECs overall. Some ATT&CK techniques are directly mapped to CAPECs, such as phishing. However, other techniques may have more than one CAPEC entry. Thus, references are not very useful for the transitive identification of threats in the organization's network.

We also studied interoperability with OWASP Top Ten list because results of scanning tools are often its categories. MITRE provides a mapping labeled as *CWE-1344* that assigns to each category its related CWEs. Appropriate counts of CWEs and CAPECs for each category are depicted in Figure 5. These counts are caused by the fact that CWEs are more granular compared to OWASP Top Ten. For example, OWASP Injection is related to SQL injection, OS command injection, and other types of injections in CWE. The categories have uneven amount of CAPECs too.
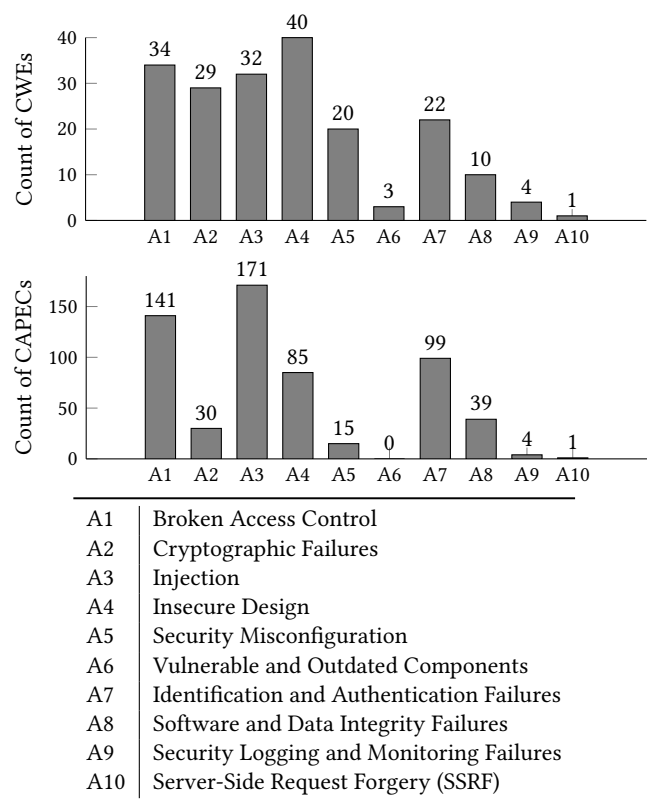
Based on these results, we conclude that mapping of enumeration entries may not accurately determine related parts of the information based on enumeration entries observed in the organization. References between enumerations from Figure 1 are ambiguous. The enumerations often contain free-text fields that enforce text-based approaches for their processing (e.g., using NLP). The text-based approach using Term Frequency – Inverse Document Frequency (TF-IDF) together with graph-based and recommendation-based approaches were provided by *Dang and François* [13].

6 CONCLUSION

In this paper, we surveyed cyber threat and vulnerability identification using enumerations about vulnerabilities and attack techniques. We revealed the current challenges and possible research directions suitable for future work. Last, we analyzed data from MITRE ATT&CK and enumerations using data sources of ATT&CK techniques and interoperability of enumeration using references between their entries.

In general, current vulnerability identification methods can determine CPE and CVE identifiers in practice. The challenges are related to the vulnerability discovery precision, a large amount of data, hybrid IT environments, and conformance to specifications. The threat identification copes, e.g., with the unstructured text of CTI reports, the overwhelming amount of data, and related lack of visibility. Enumeration entries allow the development of mature methods based on the Detection Maturity Level (DML) model.

According to the analytical results, ATT&CK provides capabilities for complex threat modeling when only network monitoring is used as a source of data. However, references between enumerations may not allow inferring related parts of the information

A1	Broken Access Control
A2	Cryptographic Failures
A3	Injection
A4	Insecure Design
A5	Security Misconfiguration
A6	Vulnerable and Outdated Components
A7	Identification and Authentication Failures
A8	Software and Data Integrity Failures
A9	Security Logging and Monitoring Failures
A10	Server-Side Request Forgery (SSRF)

Figure 5: Count of CWEs and CAPECs for OWASP Top Ten 2021 (15th January 2022).

about possible cyber threats in the organization. Scripts that allow reproducing results are provided in supplementary materials in the ACM Digital Library and on Zenodo [41].

REFERENCES

[1] David Bianco. 2013. The Pyramid of Pain. Retrieved Jan 29, 2022 from http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html

[2] Deborah J. Bodeau, Catherine D. McCollum, and David B. Fox. 2018. *Cyber Threat Modeling: Survey, Assessment, and Representative Framework.* Technical Report. The MITRE Corporation, McLean, Virginia. 1–111 pages. Retrieved Jan 29, 2022 from https://apps.dtic.mil/sti/pdfs/AD1108051.pdf

[3] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. 2014. Cyber scanning: A comprehensive survey. *IEEE Communications Surveys and Tutorials* 16 (2014), 1496–1519. Issue 3. https://doi.org/10.1109/SURV.2013.102913.00020

[4] Rebekah Brown and Robert M. Lee. 2019. *The Evolution of Cyber Threat Intelligence (CTI): 2019 SANS CTI Survey.* SANS Institute. Retrieved Jan 29, 2022 from https://www.sans.org/white-papers/38790/

[5] Brant A. Cheikes, David Waltermire, and Karen Scarfone. 2011. *Common platform enumeration: Naming specification version 2.3.* National Institute of Standards and Technology. Retrieved Jan 29, 2022 from https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir7695.pdf

[6] The MITRE Corporation. 1999-2022. CVE - Common Vulnerabilities and Exposures (CVE). Retrieved Jan 29, 2022 from https://cve.mitre.org/

[7] The MITRE Corporation. 2006-2022. CWE - Common Weakness Enumeration. Retrieved Jan 29, 2022 from https://cwe.mitre.org/

[8] The MITRE Corporation. 2007-2022. CAPEC - Common Attack Pattern Enumeration and Classification (CAPEC). Retrieved Jan 29, 2022 from https://capec.mitre.org/

[9] The MITRE Corporation. 2015-2022. MITRE ATT&CK. Retrieved Jan 29, 2022 from https://attack.mitre.org/

[10] The MITRE Corporation. 2022. Github - mitre/cti: Cyber Threat Intelligence Repository expressed in STIX 2.0. Retrieved Mar 22, 2022 from https://github.com/mitre/cti

[11] AT&T Cybersecurity. 2022. Threat Detection & Response. Retrieved Feb 1, 2022 from https://cybersecurity.att.com/solutions/threat-detection

[12] Luc Dandurand, Aaron Kaplan, Pavel Kácha, Youki Kadobayashi, Andrew Kompanek, Tomás Lima, Thomas Millar, Jose Nazario, Richard Perlotto, and Wes Young. 2014. Standards and tools for exchange and processing of actionable information. Technical Report. European Union Agency for Network and Information Security (ENISA), Heraklion, Greece. https://doi.org/10.2824/37776

[13] Quang Vinh Dang and Jerome François. 2018. Utilizing attack enumerations to study SDN/NFV vulnerabilities. In 2018 4th IEEE Conference on Network Softwarization and Workshops, NetSoft 2018. IEEE, New York, NY, USA, 329–331. https://doi.org/10.1109/NETSOFT.2018.8459961

[14] Edgescan. 2020. The Year 2020 Vulnerability statistics report. Retrieved Jan 29, 2022 from https://info.edgescan.com/vulnerability-stats

[15] Gartner, Inc. 2022. Security Orchestration, Automation and Response (SOAR). Retrieved May 21, 2022 from https://www.gartner.com/en/information-technology/glossary/security-orchestration-automation-response-soar

[16] Marian Gawron, Feng Cheng, and Christoph Meinel. 2017. PVD: Passive vulnerability detection. In 2017 8th International Conference on Information and Communication Systems (ICICS). IEEE, New York, NY, USA, 322–327. https://doi.org/10.1109/IACS.2017.7921992

[17] Béla Genge and Călin Enăchescu. 2016. ShoVAT: Shodan-based vulnerability assessment tool for Internet-facing services. Security and Communication Networks 9, 15 (10 2016), 2696–2714. https://doi.org/10.1002/sec.1262

[18] Rick Hofstede, Pavel Čeleda, Brian Trammell, Idilio Drago, Ramin Sadre, Anna Sperotto, and Aiko Pras. 2014. Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX. IEEE Communications Surveys and Tutorials 16, 4 (2014), 2037–2064. https://doi.org/10.1109/COMST.2014.2321898

[19] Ghaith Husari, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. 2017. TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of CTI Sources. In Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017). ACM, New York, NY, USA, 103–115. https://doi.org/10.1145/3134600.3134646

[20] Eric M. Hutchins, Michael J. Cloppert, and Rohan M. Amin. 2011. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research 1, 1 (2011), 80–106.

[21] IBM. 2022. IBM Security QRadar SIEM - Features. Retrieved Feb 2, 2022 from https://www.ibm.com/qradar/security-qradar-siem/features

[22] Nektaria Kaloudi and Jingyue Li. 2020. The AI-Based Cyber Threat Landscape: A Survey. ACM Computing Surveys (CSUR) 53, 1 (2020), 1–34. https://doi.org/10.1145/3372823

[23] Martin Laštovička, Martin Husák, and Lukáš Sadlek. 2020. Network Monitoring and Enumerating Vulnerabilities in Large Heterogeneous Networks. In NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium. IEEE, New York, NY, USA, 1–6. https://doi.org/10.1109/NOMS47738.2020.9110394

[24] Triet H. M. Le, Huaming Chen, and M. Ali Babar. 2021. A Survey on Data-driven Software Vulnerability Assessment and Prioritization. arXiv. https://doi.org/10.48550/ARXIV.2107.08364

[25] Gordon Fyodor Lyon. 2009. Nmap Network Scanning: The Official Nmap Project Guide to Network Discovery and Security Scanning. Insecure, USA. https://nmap.org/book/

[26] Vasileios Mavroeidis and Siri Bromander. 2017. Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence. In 2017 European Intelligence and Security Informatics Conference (EISIC). IEEE, New York, NY, USA, 91–98. https://doi.org/10.1109/EISIC.2017.20

[27] Vasileios Mavroeidis and Audun Jøsang. 2018. Data-Driven Threat Hunting Using Sysmon. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy (ICCSP 2018). ACM, New York, NY, USA, 82–88. https://doi.org/10.1145/3199478.3199490

[28] Sadegh M. Milajerdi, Rigel Gjomemo, Birhanu Eshete, R. Sekar, and V. N. Venkatakrishnan. 2019. HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows. In 2019 IEEE Symposium on Security and Privacy (SP). IEEE, New York, NY, USA, 1137–1152. https://doi.org/10.1109/SP.2019.00026

[29] Joseph Muniz, Gary McIntyre, and Nadhem AlFardan. 2016. Security Operations Center: Building, Operating, and Maintaining Your SOC. Cisco Press, Indianapolis (USA).

[30] Umara Noor, Zahid Anwar, Asad Waqar Malik, Sharifullah Khan, and Shahzad Saleem. 2019. A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories. Future Generation Computer Systems 95 (6 2019), 467–487. https://doi.org/10.1016/j.future.2019.01.022

[31] National Institute of Standards and Technology (NIST). 2018. Framework for Improving Critical Infrastructure Cybersecurity. Retrieved Jan 29, 2022 from https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

[32] The National Institute of Standards and Technology. 2022. NVD - General. Retrieved May 19, 2022 from https://nvd.nist.gov/general

[33] The National Institute of Standards and Technology. 2022. Official Common Platform Enumeration (CPE) Dictionary Statistics. Retrieved Jan 29, 2022 from https://nvd.nist.gov/products/cpe/statistics

[34] Jamie O'Hare, Rich Macfarlane, and Owen Lo. 2019. Identifying Vulnerabilities Using Internet-Wide Scanning Data. In 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3). IEEE, New York, NY, USA, 1–10. https://doi.org/10.1109/ICGS3.2019.8688018

[35] Rich Piazza, John Wunder, and Bret Jordan. 2017. STIX™ Version 2.0. Part 1: STIX Core Concepts. Retrieved Jan 29, 2022 from http://docs.oasis-open.org/cti/stix/v2.0/stix-v2.0-part1-stix-core.html

[36] Ponemon Institute, LLC. 2019. Costs and consequences of gaps in vulnerability response. Retrieved Jan 29, 2022 from https://www.servicenow.com/lpayr/ponemon-vulnerability-survey.html

[37] Sara Qamar, Zahid Anwar, Mohammad Ashiqur Rahman, Ehab Al-Shaer, and Bei-Tseng Chu. 2017. Data-driven analytics for cyber-threat intelligence and information sharing. Computers & Security 67 (6 2017), 35–58. https://doi.org/10.1016/j.cose.2017.02.005

[38] Qualys, Inc. 2022. Vulnerability Management. Retrieved Feb 2, 2022 from https://www.qualys.com/apps/vulnerability-management/

[39] Md Rayhanur Rahman, Rezvan Mahdavi-Hezaveh, and Laurie Williams. 2021. What are the attackers doing now? Automating cyber threat intelligence extraction from text on pace with the changing threat landscape: A survey. https://doi.org/10.48550/ARXIV.2109.06808

[40] Rapid7. 2022. Metasploit Documentation. Retrieved Jan 28, 2022 from https://docs.rapid7.com/metasploit/

[41] Lukáš Sadlek, Pavel Čeleda, and Daniel Tovarňák. 2022. Supplementary Materials: Current Challenges of Cyber Threat and Vulnerability Identification Using Public Enumerations. Zenodo. Retrieved Jun 20, 2022 from https://doi.org/10.5281/zenodo.6659657

[42] Sagar Samtani, Shuo Yu, Hongyi Zhu, Mark Patton, and Hsinchun Chen. 2016. Identifying SCADA vulnerabilities using passive and active vulnerability assessment techniques. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, New York, NY, USA, 25–30. https://doi.org/10.1109/ISI.2016.7745438

[43] Clemens Sauerwein, Irdin Pekaric, Michael Felderer, and Ruth Breu. 2019. An analysis and classification of public information security data sources used in research and practice. Computers & Security 82 (5 2019), 140–155. https://doi.org/10.1016/j.cose.2018.12.011

[44] Offensive Security. 2022. Exploit Database. Retrieved Jan 28, 2022 from https://www.exploit-db.com/

[45] Ryan Stillions. 2014. The DML model. Retrieved Jan 29, 2022 from http://ryanstillions.blogspot.com/2014/04/the-dml-model_21.html

[46] Tenable, Inc. 2022. Nessus. Retrieved Jan 28, 2022 from https://www.tenable.com/products/nessus

[47] Tenable, Inc. 2022. Tenable.io vulnerability management. Retrieved Feb 2, 2022 from https://www.tenable.com/products/tenable-io

[48] Wiem Tounsi and Helmi Rais. 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & Security 72 (1 2018), 212–233. https://doi.org/10.1016/j.cose.2017.09.001

[49] Daniel Tovarňák, Lukáš Sadlek, and Pavel Čeleda. 2021. Graph-Based CPE Matching for Identification of Vulnerable Asset Configurations. In 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM). IEEE, New York, NY, USA, 986–991. https://ieeexplore.ieee.org/abstract/document/9463994

[50] Thomas D. Wagner, Khaled Mahbub, Esther Palomar, and Ali E. Abdallah. 2019. Cyber threat intelligence sharing: Survey and research directions. Computers & Security 87 (11 2019), 101589. https://doi.org/10.1016/j.cose.2019.101589

[51] Suzanne Widup, Dave Hylender, Gabriel Bassett, Philippe Langlois, and Alex Pinto. 2020. 2020 Data Breach Investigations Report. Technical Report. Verizon. https://doi.org/10.13140/RG.2.2.21300.48008

[52] Wenjun Xiong and Robert Lagerström. 2019. Threat modeling – A systematic literature review. Computers & Security 84 (7 2019), 53–69. https://doi.org/10.1016/j.cose.2019.03.010

[53] Jun Zhao, Qiben Yan, Jianxin Li, Minglai Shao, Zuti He, and Bo Li. 2020. TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. Computers & Security 95 (8 2020), 101867. https://doi.org/10.1016/j.cose.2020.101867