How does Endpoint Detection use the MITRE ATT&CK Framework?

Apurva Virkud, Muhammad Adil Inam, Andy Riddle, Jason Liu, Gang Wang, and Adam Bates, *University of Illinois Urbana-Champaign*

https://www.usenix.org/conference/usenixsecurity24/presentation/virkud

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

How does Endpoint Detection use the MITRE ATT&CK Framework?

Apurva Virkud, Muhammad Adil Inam, Andy Riddle, Jason Liu, Gang Wang, Adam Bates
University of Illinois Urbana-Champaign
{avirkud2, mainam2, rriddle2, jdliu2, gangw, batesa}@illinois.edu

Abstract

MITRE ATT&CK is an open-source taxonomy of adversary tactics, techniques, and procedures based on real-world observations. Increasingly, organizations leverage ATT&CK technique "coverage" as the basis for evaluating their security posture, while Endpoint Detection and Response (EDR) and Security Indicator and Event Management (SIEM) products integrate ATT&CK into their design as well as marketing. However, the extent to which ATT&CK coverage is suitable to serve as a security metric remains unclear— *Does ATT&CK coverage vary meaningfully across different products? Is it possible to achieve total coverage of ATT&CK? Do endpoint products that detect the same attack behaviors even claim to cover the same ATT&CK techniques?*

In this work, we attempt to answer these questions by conducting a comprehensive (and, to our knowledge, the first) analysis of endpoint detection products' use of MITRE ATT&CK. We begin by evaluating 3 ATT&CK-annotated detection rulesets from major commercial providers (Carbon Black, Splunk, Elastic) and a crowdsourced ruleset (Sigma) to identify commonalities and underutilized regions of the ATT&CK matrix. We continue by performing a qualitative analysis of unimplemented ATT&CK techniques to determine their feasibility as detection rules. Finally, we perform a consistency analysis of ATT&CK labeling by examining 37 specific threat entities for which at least 2 products include specific detection rules. Combined, our findings highlight the limitations of overdepending on ATT&CK coverage when evaluating security posture; most notably, many techniques are unrealizable as detection rules, and coverage of an ATT&CK technique does not consistently imply coverage of the same real-world threats.

1 Introduction

It is difficult to overstate the influence of the MITRE Corporation's ATT&CK knowledge base [79] on how we conceptualize today's threat landscape. ATT&CK catalogues the observed real-world behaviors (*Procedures*) of hundreds of sophisticated threat groups. It then systematizes these procedures by assigning them to a known adversarial *Technique*, an explanation of "how" the attacker is attempting to achieve an operational goal. Techniques are themselves grouped into *Tactics* that explain the "why" of that goal, supplanting prior notions of a cyber "kill chain" [35]. The end result is a hierarchical taxonomy that bridges the gap between high-level attack objectives and concrete activities on targeted systems.

MITRE ATT&CK is invaluable as a means of systematizing seemingly-disparate attack behaviors and provides much needed context for threat alerts. The US Cybersecurity & Infrastructure Security Agency advises the use of mapping raw data to ATT&CK techniques as a means of enriching threat intelligence and cybersecurity advisories [17]. By annotating low-level threat intelligence with ATT&CK, experts and non-experts alike can situate a threat in the broader context of possible adversary actions and objectives. Today's threat detection products, most notably EDR's and SIEM's, annotate nearly every threat detection rule with ATT&CK techniques (e.g., [21,78,86]). Thus, when an alert occurs, analysts benefit from not only a specific description of the detection query, but also an explanation for how the event fits into the bigger picture of an attack pattern.

However, the applications of MITRE ATT&CK go beyond systematization and explainability – detection products (and the organizations that employ them) are now regularly evaluated on their ability to "cover" each of the ATT&CK techniques. A 2020 survey of security professionals finds that 57% of respondents also use ATT&CK to evaluate the efficacy of deployed security products [9]. Digital forensics and incident response consultants now regularly conduct audits of their clients' coverage of the ATT&CK framework (e.g., [65]). Soon, organizations' cyber insurance premiums may even consider ATT&CK coverage [5]. Unsurprisingly, it is also now commonplace for vendors to actively tout their coverage of ATT&CK in marketing materials (e.g., [15, 31, 70, 87]). However, despite its importance, little attention has been paid to how endpoint detection products actually employ MITRE

ATT&CK, or whether the hype around ATT&CK coverage as a security metric is justified.

In this work, we conduct an independent analysis of MITRE ATT&CK's use in endpoint detection products. *We emphasize that our intent is to evaluate the suitability of MITRE ATT&CK coverage as a security metric, rather than the products themselves.* Analyzing the rule sets of 3 major products, we answer the following research questions:

RQ1: How do products use ATT&CK? We conduct an empirical analysis of how endpoint detection rules are annotated with ATT&CK and examine the overall ATT&CK coverage of popular products. We find that products do not attempt to cover all ATT&CK techniques, with coverage ranging from 48% to 55%. Further, we observe that the available level of coverage is inflated by the presence of low risk and low severity rules that are less likely to be prioritized in practice. Filtering out low and medium risk rules, technique coverage drops to 25%–26%, approximately half of the original coverage. Finally, in addition to the total coverage of techniques between products being fairly consistent, we find that the products have similar preferences for which ATT&CK techniques to cover with statistical significance.

RQ2: Why don't products detect all of ATT&CK? Many ATT&CK techniques (53, 27.7%) were not implemented in any of the commercial products. To understand why, three authors performed open coding on the descriptions of these unimplemented techniques and come to a consensus for all codes. The identified reasons are (1) ineffective detection methods, many of which MITRE explicitly mentions will exhibit high false positive rates and other difficulties; (2) unsuitable target infrastructure, such as techniques that target non-host systems or social media; and (3) techniques that require knowledge of the client enterprise environment.

RQ3: How consistently is ATT&CK applied? Finally, we examine how consistently different vendors perform ATT&CK labeling when attempting to detect the same threats. For this comparison, we identify malicious entities (e.g., malware, CVEs, threat actors) across the rule sets by matching rule metadata against a known list of malicious entities from the rule descriptions. Examining 37 malicious entities that are explicitly referenced in at least two rulesets, we find that vendors are applying ATT&CK technique labels in equally-valid but inconsistent ways. We identify cases of rules that overlap in detected behavior but differ in the annotated ATT&CK techniques and tactics, such that a security analyst may reach different conclusions about the same threats depending on which product they are using.

In addition to three commercial products, we augment our analysis by replicating RQ1 for a large *crowdsourced* ruleset. As crowdsourced rules may not be subject to a uniform quality control process, we report these results in Appendix A. While the coverage of MITRE ATT&CK techniques (79%) is higher than the commercial rulesets, we observe similar trends in the rankings of covered techniques (with statistical significance) as well as the presence of low-criticality rules that may artificially inflate ATT&CK coverage.

To our knowledge, this work marks the first independent analysis of how MITRE ATT&CK is integrated into real-world endpoint detection products. We emphasize that the coverage of ATT&CK is already used as a metric for assessing these types of systems and thus has impacts on real world security. We conclude by discussing the implications of our findings for the enterprise security ecosystem at large. We find that ATT&CK, while useful for explanation purposes, is a poor measure of the detection capabilities of an endpoint detection product. This observation was supported by discussions with one of the surveyed product vendors during our disclosure process. In light of this, we advise stakeholders in the security ecosystem to approach coverage-based evaluations of organizations and products with caution and nuance. Finally, we advise MITRE to take a more active role in guiding how the ATT&CK framework is employed to mitigate future misuse. We open-source our code and analysis at [84] to support future work in the area.

2 Background

MITRE ATT&CK. MITRE ATT&CK is a hierarchical knowledge base that systematizes real-world observations of adversary *procedures* into general *techniques* and high-level *tactics* [79]. Techniques classify an observed procedure from a given threat group into a common offensively-oriented action (the "how"), while Tactics identify the associated attacker objective(s) (the "why"). Procedures can be associated with one or more techniques, and techniques can be associated with one or more tactics. Techniques can also be divided into sub-techniques that describe more granular behavior.[1] As of Version 11, the enterprise ATT&CK matrix that we use in this work is comprised of 14 tactics and 191 techniques. While best known for its taxonomy of tactics and techniques, ATT&CK also indexes additional information including verified threat groups, adversarial campaigns, detection/mitigation strategies, telemetry data sources, and software.

Endpoint Detection Products. Today's enterprises employ a variety of products for threat detection and remediation [39]. Endpoint Detection and Response (EDR) systems are a critical component of enterprise security [24, 26]. EDRs capture and examine endpoint telemetry data for evidence of potentially malicious activity. Security Indicator and Event Management (SIEM) software (e.g., [78]) ingest a variety of telemetry streams, including endpoint events, to centralize analyst operations. SIEMs also typically provide additional detection capabilities over endpoint events, making them functionally equivalent to EDRs for the purposes of this work.

While machine learning (ML) is increasingly integrated

[1] We do not consider sub-techniques in this work because it was not common practice for the surveyed EDRs to annotate rules with sub-techniques.

into products, endpoint detection functions are still largely *rule-based* (i.e., *heuristic-based*) [24, 26], in which an analyst explicitly defines a search query that describes a known adversary behavior. Rule-based detection, as opposed to anomaly-based or other ML approaches, enjoys the advantages of being fully *explainable* because every detection rule is annotated with metadata that explains the intended detection behavior. Today, this metadata often includes annotations linking a rule to one or more techniques enumerated by the MITRE ATT&CK framework. Typically, security analysts will annotate rules with ATT&CK techniques manually based on individual judgement. Rule-based detection also allows operators to tune the detection behavior on a rule-by-rule basis; if a rule is causing many false alarms in an organization, it can be disabled or deprioritized. Products also provide investigation features that allow analysts to triage alerts and examine the telemetry that caused them to fire, as well as response features such as automatic quarantine or malicious process removal.

Detection Rules. To further understand how detection rules work, we demonstrate the syntax and general detection strategy of three exemplar endpoint detection products – Carbon Black [86], Splunk [78], and Elastic [21].

```
process_name:wevtutil.exe
and process_cmdline:cl*
and -process_cmdline:clicktorun*
and -process_cmdline:AnyConnect\.evtx*
```

This Carbon Black rule, obtained from [1], searches for evidence that an attacker is using the Windows Event Utility `wevtutil.exe` to destroy application or system logs. The rule also checks that certain commandline strings are *not* present with the negation operator to tune out common legitimate use cases where an administrator is clearing logs. The rule is tagged with the ATT&CK technique `T1070`, Indicator Removal, which links to the Defense Evasion `TA0005` tactic.

```
(Processes.process_name="RDPWInst.exe"
  OR Processes.original_file_name= "RDPWInst.exe")
AND Processes.process IN ("* -i*", "* -s*",
    "* -o*", "* -w*", "* -r*")
```

This Splunk rule, obtained from [78], searches for `RDPWInst.exe`, which is a Remote Desktop Protocol wrapper library tool that can be abused for remote access. The rule is tagged with the ATT&CK technique `T1021`, Remote Services, which links to the Defense Evasion `TA0005` tactic.

```
event.category : (network or network_traffic)
and network.transport:tcp
and (destination.port: 26
     or (event.dataset:zeek.smtp
         and destination.port: 26))
```

This Elastic rule, obtained from [21], searches for Simple Mail Transfer Protocol (SMTP) traffic on the (non-default) port 26. While legitimate mail transfer agents may use port 26 to deconflict with other agents, it is also used by the BatPatch malware family for command and control traffic. This rule is tagged with the ATT&CK technique `T1048`, Exfiltration Over Alternative Protocol, which links to the Command and Control (`TA0011`) and Exfiltration (`TA0010`) tactics.

Stats	CB	Splunk	Elastic	Sigma
# ATT&CK Tagged Rules	867	911	473	2,195
# Unique Techniques	105	100	92	151
% Technique Coverage	55%	52%	48%	79%
Tactic Coverage	13/14	14/14	13/14	14/14

Table 1: **Dataset Overview:** Detection rules annotated with MITRE ATT&CK technique labels in each product out of all rules present in the head of the main branch of the rule repository (or, for Carbon Black, the customer-visible watchlists).

Marketing with ATT&CK. Beyond the product, MITRE ATT&CK is used extensively in marketing. We identify no less than 19 security vendors that have appealed to ATT&CK in advertisements [12–16, 18, 31, 34, 36, 37, 42, 45, 50, 53, 57, 62, 70, 73, 74]. Notably, 12 of these sources are specifically advertising performance in the MITRE Engenuity sponsored ATT&CK Evaluations. While MITRE Engenuity's guidelines recommend investigating low-level detection details to determine a system's performance, these promotional materials report high-level metrics like ATT&CK coverage. In this work, we investigate the validity of using ATT&CK coverage as a measure of product efficacy.

3 Dataset

We collect data from four popular rule engines and analyze how they make use of the MITRE ATT&CK framework to cover different attack tactics and techniques. We consider detection rules from three popular industry endpoint detection systems: VMware Carbon Black (CB) [86], Splunk Security Content [78], and Elastic [21]. Splunk and Elastic's rulesets are fully open-sourced, available on GitHub. Carbon Black shares hundreds of rules with their customers in the form of curated watchlists that can be enabled for their deployment. We received access to the Carbon Black Cloud product via an Educational license and received their permission to use their name and ruleset in this paper. Carbon Black was reported to be one of the top EDR solutions in multiple 2023 market reports [24,26], while both Splunk and Elastic are popular SIEM products [59]. Additionally, we include an open and crowd sourced rule repository curated by Sigma [72], also available on GitHub. The ruleset uses a vendor agnostic format and has downstream users including other industry systems such as IBM QRadar. We select these systems due to their wide deployment. Because a crowdsourced ruleset may not have uniform processes to verify the quality of rules, we report on Sigma separately in Appendix A. Our intention is not to evaluate the quality of individual products, but instead to use these products to gain insight into how MITRE ATT&CK is integrated into endpoint detection.

Table 1 summarizes the key statistics of the dataset, which was captured by taking a snapshot of each ruleset in October

Data Field	CB	Splunk	Elastic	Sigma
Name of Attack	✓	✓	✓	✓
Description	✓	✓	✓	✓
ATT&CK Technique(s)	✓	✓	✓	✓
Known False Positives	✓*	✓	✓	✓
Confidence	✓*	✓		
Risk Score		✓	✓	
Severity Score	✓		✓	
Keywords	✓		✓	
References		✓	✓	✓

Table 2: **Metadata from Carbon Black (CB), Splunk, Elastic, and Sigma:** Metadata fields that appear in more than one ruleset, indicated by a checkmark. The asterisk (*) indicates that the field was extrapolated from another field.

2022. We filtered a single crowdsourced watchlist containing 68 rules from the Carbon Black ruleset because these may not undergo quality verification from Carbon Black itself. We also filtered rules that were not annotated with ATT&CK techniques. Finally, we omit rules from the three commercial systems (all except Sigma) that are marked as deprecated or in development, as either indicates that they are not currently endorsed by the product vendor. We refrain from describing statistics of the filtered rules in detail so as to avoid benchmarking the rulesets of the different systems; this is not our objective. Broadly, the majority of rules from all four products are annotated with ATT&CK technique tags, with the exception being rules that are simple IP blocklist rules (e.g., Tor exit nodes). In total, we identify 867 Carbon Black, 911 Splunk, 473 Elastic, and 2,195 Sigma rules.

Our analysis is also informed by the rule metadata, shown in Table 2, made available by each vendor. While all four rulesets include a name, description, and ATT&CK technique tags for each rule, fields begin to diverge subtly beyond this basic information. Splunk, Elastic, and Sigma all contain explicit fields to denote known sources of false positives for each rule, while Carbon Black sorts its rules into recommendation lists based in part on the likelihood of false positives for a given rule. Carbon Black also implicitly describes confidence[2] in these list descriptions, while Splunk has an explicit score, and Elastic has no confidence indicator at all. Only Splunk and Elastic have risk scores[3] while only Carbon Black and Elastic have severity scores,[4] etc.

4 How do products use ATT&CK?

Using this dataset, we analyze how endpoint detection products use the MITRE ATT&CK framework. In this section we will focus on the three commercial rulesets—Carbon Black, Splunk, and Elastic. The corresponding results on the crowdsourced ruleset Sigma can be found in Appendix A.

More specifically, we are interested in understanding which ATT&CK techniques or tactics have corresponding rules implemented, and the coverage of the rulesets. As of Version 11, ATT&CK contains 14 *tactics* that describe the high-level goals of the different phases of the attack and 191 *techniques* that describe the specific attacker actions and methods under these tactics. As shown in Table 1, while all three rulesets cover the vast majority of the tactics (at least 13/14), their overall technique coverage is 48%–55%. In particular, there are 53 techniques (27.7%) that do not have a corresponding rule in any of the three rulesets. Our analysis below shows that different tactics and techniques receive uneven attention or coverage across different products. Certain techniques are consistently under-covered by all three rule engines.

ATT&CK Technique Coverage. Figure 1 reports technique

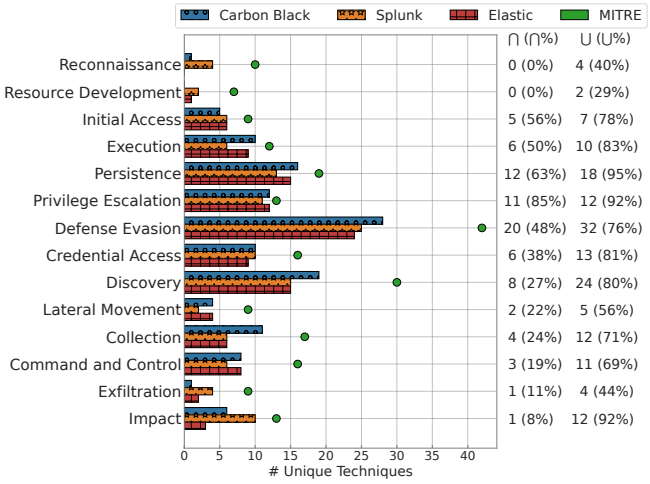

Figure 1: **Technique Coverage under Each Tactic:** The y-axis shows the 14 ATT&CK tactics ordered by the phase of attack (e.g., reconnaissance is typically the first step). The green dots represent the total number of techniques under each tactic in ATT&CK and the bars represent the number of covered techniques by each ruleset. The number and % of covered techniques in the intersection (∩) and union (∪) of all rulesets are shown on the right side.

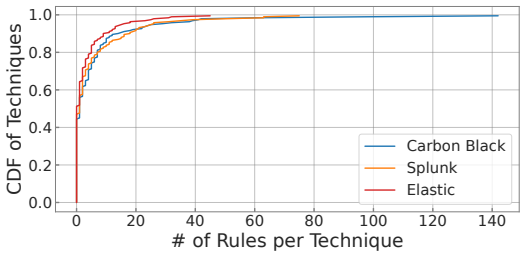

Figure 2: **Rules Per Technique:** Cumulative distribution of the number of rules per technique for each endpoint detection system.

[2]Confidence is the likelihood that an alert is indicative of an attack.
[3]Risk scores are a composite of confidence and severity measures.
[4]Severity scores reflect the potential damage should the attack occur.

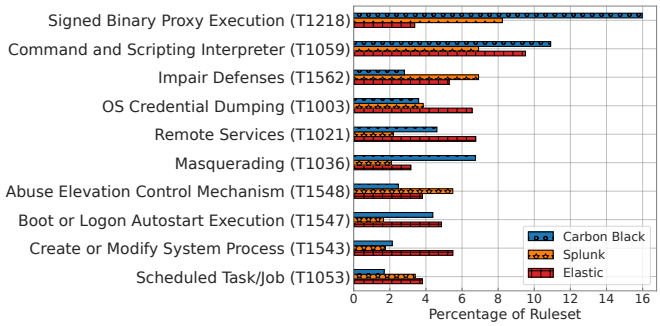

Figure 3: **Top ATT&CK Techniques:** Top 10 techniques ranked based on the sum of associated rules from the three engines. The x-axis shows the % of the rules for a given technique in each engine (e.g., 16% of Carbon Black rules are tagged with `T1218`).

coverage, by tactic, for the three products. The green dot denotes the total number of ATT&CK techniques under each tactic and the bars represent the number of techniques with associated rules in each product. Interestingly, we observe that tactics with more ATT&CK techniques also receive more attention/coverage from the product. More specifically, *defense evasion*, *discovery*, and *persistence* are among the most frequently appearing tactics across all three products. They are also the tactics with the highest number of techniques in the ATT&CK framework. However, certain tactics are consistently under-covered by all three engines. For example, *resource development* only has 2 techniques covered (no coverage by Carbon Black) and *reconnaissance* only has 4 techniques covered (no coverage by Elastic). We suspect these techniques describe offline activities for which it is difficult to implement rules, a hypothesis we explore further in §5.

In Figure 1, we also report the intersection ($\cap$) and union ($\cup$) of the techniques covered by the three products. We observe that the intersection on average covers 32% of the techniques under a given tactic, while the union of the three engines can boost the coverage on average by 38%. The exception is *privilege escalation* where the intersection coverage is already high (85%) and the union can only boost the coverage to 92%. This kind of multi-product analysis of MITRE ATT&CK coverage forms the basis of org-level evaluations of security posture (e.g., [65]); a security consultant evaluating an organization that only licensed one of these products might argue that the organization could improve its security by licensing another of the three products.

ATT&CK Technique Density. We also observe differences not only in the coverage of techniques, but also in their frequency, as shown in Figure 2. We see that 9.9%-13.1% of techniques only have one associated rule, while 29.8%-32.5% techniques have 1 to 5 rules. A small fraction of techniques (1.6%-2%) have more than 50 associated rules implemented. An extreme example is `T1218` (System Binary Proxy Execution) for which Carbon Black has implemented 142 rules to detect this threat. The general trend is consistent across the three products.

Metric	Filter	Carbon Black	Splunk	Elastic
Baseline	No Filter	55%	52%	48%
Risk	>= Med.	/	43%	42%
	>= High	/	25%	26%
Severity	>= Med.	52%	/	42%
	>= High	46%	/	26%
Confidence	>= Med.	/	51%	/
	>= High	/	46%	/

Table 3: **Impact of Risk/Severity/Confidence on ATT&CK Technique Coverage:** We observe that technique coverage drops drastically when only considering the rules with the highest operational value. "/" means metadata is not available.

Figure 3 lists the top 10 techniques aggregated across all three rulesets. While all 10 techniques have a high coverage by the three engines, there is some variation per ruleset. For example, `T1218` (Signed Binary Proxy Execution) is the predominant technique in Carbon Black and Splunk, but it is not the most frequent in Elastic. To further measure the per-ruleset variation, we calcuate the Spearman's rank correlation coefficient [20, 75] between the ranked lists of techniques for each pair of rulesets. The Spearman coefficient is calculated as $\rho_{R(A),R(B)} = \frac{cov(R(A),R(B))}{\sigma_{R(A)}\sigma_{R(B)}}$ where lists of covered techniques A and B have rank variable representations $R(A)$ and $R(B)$ (in this case, techniques are mapped to the percentage of the ruleset they cover), *cov* is the covariance, and σ is the standard deviation. The coefficient value ρ ranges from -1 to 1 and a positive value closer to 1 indicates more similar ranking for our variables. We also conduct a t-test to determine if ρ is significantly different than 0.

Interestingly, we observe a high level of consistency among the three products in terms of techniques with implemented rules. More specifically, the Spearman coefficient between any given pair of ranked lists is always positive: 0.634 for Carbon Black and Splunk, 0.744 for Carbon Black and Elastic, and 0.639 for Splunk and Elastic. The p-value is < 0.001 for all tests, indicating a statistically significant similarity between the technique rankings of each pair of products. Collectively, it means that, even though different ATT&CK techniques have received uneven attention, the three engines have a similar preference in terms of which techniques to cover.

Risk, Severity, Confidence, and their Impact on ATT&CK Coverage. To provide further context for the implemented rules, we analyze the quantitative metrics assigned to the rules, including confidence, risk, and severity scores. These metrics are supposed to help security analysts to triage alerts. We note that these metrics are assessments made by the developers who work on the rules and are not necessarily universal across endpoint products. However, they serve as an approximation for how the developers believe the rule should be used in

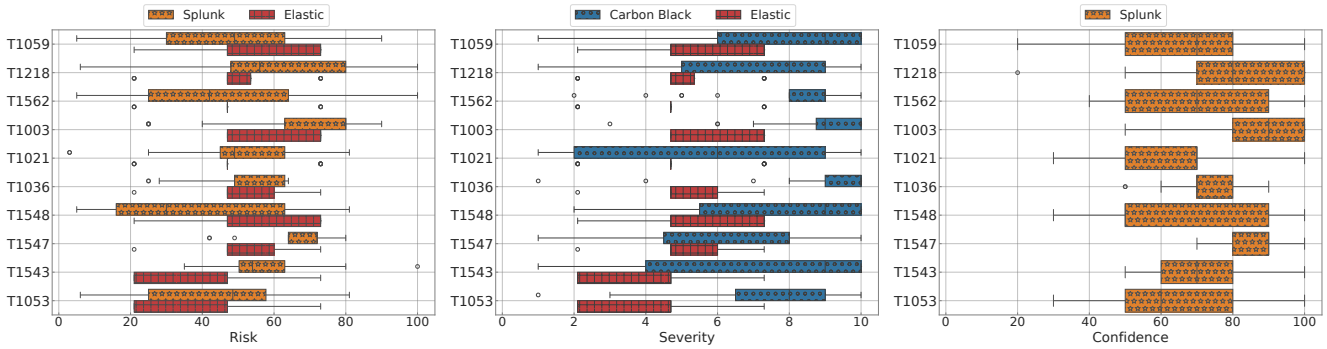

Figure 4: **Risk, Severity, and Confidence For Top Techniques:** The risk, severity, and confidence scores are not available in all three rulesets (see Table 1). We report the available scores for the top 10 techniques ranked by the number of associated rules (see Figure 3).

Metric	Filter	Carbon Black	Splunk	Elastic
Baseline	No Filter	13	14	13
Risk	>= Med.	/	14	12
	>= High	/	13	12
Severity	>= Med.	12	/	12
	>= High	12	/	12
Confidence	>= Med.	/	14	/
	>= High	/	14	/

Table 4: **Impact of Risk/Severity/Confidence on ATT&CK Tactic Coverage:** There are in total 14 tactics in the framework. "/" means metadata is not available to support the filtering.

deployment. Intuitively, rules with high values are likely to be prioritized. On the contrary, rules with low risk/severity/-confidence scores may have low operational value in practice. Splunk confirmed this intuition in their documentation on how risk scores are calculated [77], while Carbon Black confirmed to us internally that this is their intended usage. In Appendix B, we have provided some examples of rules that are likely to have false positives and have lower values in these metrics. By considering these metrics for rules of the same techniques *under the same product*, we examine how each engine interprets the risk of these techniques. For any given metric, the rulesets consistently report scores either with a range of (0, 10) or (0, 100), so we do not need to normalize the metric across products.

Figure 4 shows the risk, severity, and confidence score distributions for each of the top 10 techniques. Recall that all scores are not available in all three rulesets (see Table 1). For example, only Splunk and Elastic report the risk score.

The leftmost plot shows that, for the same technique, there is a *high variance* in terms of the risk scores for the associated rules within each engine. A possible explanation is that there is a wide spectrum of behaviors under the same ATT&CK technique that have different risk levels. The same observation applies to the center plot, reporting severity for Carbon Black and Elastic, indicating that each ATT&CK technique has a large room for different interpretations in terms of the risk/-severity even within an individual product. In other words, the technique itself does not necessarily indicate the perceived risk or severity of the attack. For confidence (rightmost plot), only Splunk has reported this score. While most of the rules have a confidence score over 50 (out of 100), their variance under individual ATT&CK techniques is also high.

Finally, we investigate how the risk, severity, and confidence metrics affect the overall MITRE ATT&CK coverage. The intuition is that rules with lower levels of these metrics have lower operational value and are unlikely to be prioritized by analysts during attack investigation. Therefore, we re-examine ATT&CK coverage after filtering out rules with lower operational values. More specifically, NIST defines the Common Vulnerability Scoring System (CVSS, v3.0) [55] which maps quantitative values (ranging from 0.0 to 10.0) to qualitative severity categories: "None", "Low", "Medium", "High", and "Critical". Considering that the risk and confidence metrics are closely related to the severity score, and all metrics have a (0, 10) or (0, 100) range, we apply the same CVSS mapping for all three metrics (as an approximation).

In Table 3 and Table 4 we report the technique coverage and tactic coverage by considering rules with at least "medium" score (or "high" score). Notably, for both risk and severity, we observe that the ATT&CK technique coverage is halved for Splunk and Elastic when only considering rules with "high" or "critical" levels (i.e., the coverage drops from 52% and 48% to 25% and 26%, respectively).

For tactics, we observe that Elastic loses coverage of "resource development" when eliminating low risk rules, while Splunk loses coverage of "reconnaissance" when eliminating low and medium risk rules. A similar effect is observed for severity: when eliminating low severity rules, Elastic loses coverage of "resource development" while Carbon Black loses coverage of "reconnaissance". This indicates that rulesets do not have effective coverage of these two earliest stages of attack development.

Rules with Multiple ATT&CK Techniques. While the vast majority of rules are annotated with a single technique (84.5%), we find that 349 rules across the three rulesets

(15.5%) have multiple technique annotations. This may be because that the rule broadly detects different system activities, or because the activity can be employed at multiple phases of attack. The breakdown of rules with multiple techniques in each ruleset can be found in Appendix Figure 10. Rules with 2 technique annotations are somewhat common, but rules with 3 or more techniques account for only 1.2%–3.1% of the rules in each ruleset. We investigate an example to understand why a rule may be annotated with multiple techniques.

```
(Processes.parent_process_name=wmiprvse.exe
  OR Processes.parent_process_name=services.exe
  OR Processes.parent_process_name=svchost.exe
  OR Processes.parent_process_name=wsmprovhost.exe
  OR Processes.parent_process_name=mmc.exe)
(Processes.process_name=powershell.exe
  OR (Processes.process_name=cmd.exe
    AND Processes.process=*powershell.exe*)
  OR Processes.process_name=pwsh.exe
  OR (Processes.process_name=cmd.exe
    AND Processes.process=*pwsh.exe*))
```

This rule from Splunk [78] is annotated with six techniques, shown in Appendix Table 7.

This rule detects parent processes that are commonly used in lateral movement behavior and spawn Powershell child processes. In this case, the multiple technique annotations are due to the breadth of the detection. Depending on which parent process is detected, Windows Management Instrumentation (T1047) or Remote Services (T1021) may apply. The detected behavior is also utilizing command line (T1059) and trusted system processes (T1218, T1543). We observe that these techniques are in total associated with five unique tactics, indicating that the system behavior may be applicable at different phases of an attack.

Sigma. We replicate the above analysis on the Sigma ruleset and briefly discuss the results. The full analysis can be found in Appendix A. First, we compare the techniques ranked by coverage in Sigma to the technique rankings in each of the other three rulesets and find that the Spearman coefficient is always moderately positive with statistical significance. Thus, despite Sigma's higher overall technique coverage (79%, see Table 1), the techniques that are covered are prioritized similarly to the commercial rulesets. The distribution of rules per technique is also similar to the commercial rulesets with a long tail of techniques that each represent less than 0.5% of the total rules. Sigma is also similar to the commercial rulesets in that filtering out low and medium criticality level rules drops the coverage of MITRE ATT&CK techniques.

5 Why don't products detect all of ATT&CK?

So far, our results show that endpoint detection products are not using all of the MITRE ATT&CK techniques to construct their detection rules, with coverage ranging from 48% to 55%. In particular, there are 53 ATT&CK techniques (27.7%) that are not implemented by any of the three commercial products. To understand *why* products may not implement corresponding rules for these techniques, we perform a *qualitative analysis* on the textual description of ATT&CK techniques.

Label	Techniques	Example
Ineffective Detection Method	21 (39.6%)	T1480
Targeting Non-Host Infrastructure	13 (24.5%)	T1584
Client-specific	9 (17.0%)	T1528
Vague Detection Method	9 (17.0%)	T1602
Targeting Third Parties	8 (15.1%)	T1591
Provenance-based Detection	4 (7.5%)	T1578
Involving Low-level Behavior	3 (5.7%)	T1200
Involving Removable Media	3 (5.7%)	T1025
Involving Human Factors	1 (1.9%)	T1598
Reason Unknown	2 (3.8%)	T1217
Total Unique Techniques	53	

Table 5: **Qualitative Labels for Unimplemented Techniques:** We label the 53 techniques that are not implemented in any of the three endpoint detection rulesets. Note that one technique may have multiple labels.

5.1 Qualitative Analysis Method

Our analysis is focused on the textual technique description as well as the listed detection strategies in the MITRE ATT&CK framework. To extract the high-level reasons (and a code book), three coders independently analyze (with open coding) the 53 ATT&CK techniques that do not have corresponding rules in any of the three commercial products. All coders are active researchers in the area of intrusion detection and are familiar with the MITRE ATT&CK framework. After independently coding the same subset of techniques, two of the coders meet and discuss their codes to decide upon common terminology and code definitions. Then these two coders continue to independently code the rest of the techniques and refine the codebook. To verify soundness of the coding results, a third coder first independently performs open coding on the techniques to confirm that no new codes emerge and then maps those codes to the existing codebook. Eventually, the three coders code all of the 53 ATT&CK techniques that are not implemented by any product. After independently coding, the coders discuss each technique to verify codes and resolve any disagreements, coming to a consensus on all codes. Since all technique codes are collaboratively reviewed by multiple researchers, we do not report inter-rater reliability [46].

5.2 Annotation Results

As shown in Table 5, we are able to attribute potential reasons behind the lack of implementation for 51 out of 53 techniques (96.2%). Note that one technique may have multiple associated reasons. We are unable to attribute reasons for two

techniques (marked as "reason unknown" in Table 5), which will be further discussed below.

Ineffective Detection Method. 21 techniques (39.6%) have ineffective detection methods, which is the most predominant reason. Among them, MITRE explicitly mentions that the suggested detections are ineffective for 14 techniques. For example, `T1480` (Execution Guardrails) refers to attackers using guardrails to only execute an attack when their desired environment conditions are fulfilled (to evade detection). MITRE ATT&CK suggests monitoring for suspicious processes and command executions that gather system information, but MITRE also specifies that this behavior can be difficult to detect since it depends on how the attacker implements their guardrail (i.e., easily producing false positives). For the remaining 7 techniques, detection methods are assessed to be ineffective by the coders. One example of such a technique is `T1594` (Search Victim-Owned Websites), which refers to attackers searching websites owned by the victim for information that can be used during targeting. The detection suggestion is to look for suspicious network traffic, which again can lead to high rates of false positives.

Targeting Non-Host Infrastructure. 13 techniques (24.5%) target non-host systems that are part of the target organization's infrastructure. For example, `T1584` (Compromise Infrastructure) refers to threats from compromised infrastructures such as cloud servers and remote repositories. The suggested detection is active Internet scanning to identify such compromises. Since the target is not the end-point machine, endpoint products may not be best suited for detection.

Client-specific. 9 techniques (17.0%) are client-specific, which are dependent on the specific services/applications on the clients' hosts. For example, `T1528` (Steal Application Access Token) depends on specific applications on hosts. Products may not implement rules for these techniques because they require knowledge of specific services or parameters that customers use.

Vague Detection Method. 9 techniques (17.0%) have detection methods that are labeled as too vague. For example, the detection for `T1602` (Data from Configuration Repository) suggests monitoring network traffic for anomalies but does not include specific heuristics. This type of detection may be more difficult to implement by an endpoint detection system.

Targeting Third Parties. 8 techniques (15.1%) involve behavior on third-party platforms that are outside of the target organization's infrastructure (e.g., open websites). For example, `T1591` (Gather Victim Org Information) involves searching an organization's social media and largely takes place outside the control of the endpoint product defenses.

Provenance-based Detection. 4 techniques (7.5%) use detection methods requiring provenance tracing, such as `T1578` (Modify Cloud Compute Infrastructure), and suggest viewing events as a chain of behavior. Such capability is often not yet available for rule-based detection systems.

Involving Low-level Behavior. 3 techniques (5.7%) involve low-level behavior (e.g., in hardware or firmware) that are not detectable by endpoint detection. An example is `T1200` (Hardware Additions).

Involving Removable Media. techniques (5.7%) involve removable media such as USBs, where an endpoint detection product may not be able to determine the correct mount path to monitor (e.g., non-`C:` drives on Windows). For example, `T1025` (Data from Removable Media) concerns data collection from such sources.

Involving Human Factors. 1 technique (1.9%) involves human factors: `T1598` (Phishing for Information). We assume that indicators resulting from human involvement would be difficult to encode in a rule.

Reasons Unknown. We were unable to determine why 2 techniques have not been implemented. For example, `T1217` (Browser Information Discovery) covers behavior that should be visible within system logs. The second case is `T1615` (Group Policy Discovery), where the detections provide examples of system-level signatures associated with abnormal active directory access. It is possible that endpoint products do not implement rules for these techniques because they are not prevalent in real-world settings or the perceived risk of such behavior is relatively low (but we are unable to confirm).

In summary, we find that many techniques are difficult (if not impossible) to implement as effective detection rules due to vague descriptions of attack behaviors or ineffective detection strategies, or because the attacker actions are outside the scope of a typical endpoint detection product. The implication is that ATT&CK coverage may not be a sound security metric since covering all these ATT&CK techniques could mean sacrificing the quality of the detection rules.

6 How consistently is ATT&CK applied?

After investigating techniques that are not implemented by products, we now focus on the *implemented rules* and examine whether products have applied MITRE ATT&CK consistently to tag the rules. As a comprehensive knowledge base, ATT&CK provides a common language for describing and communicating security threats between different vendors/-parties in the community. For rules created to detect the same threat (attacker action), we expect them to be tagged with the same ATT&CK techniques such that people can effectively link and compare rules from different vendors. Note that we are not considering the *process* of how analysts annotate rules with ATT&CK; rather, we aim to analyze the resulting annotations. For this analysis, we first search for rules that are created to detect the same malicious entities across the three products, and empirically assess the consistency of their tagged ATT&CK techniques.

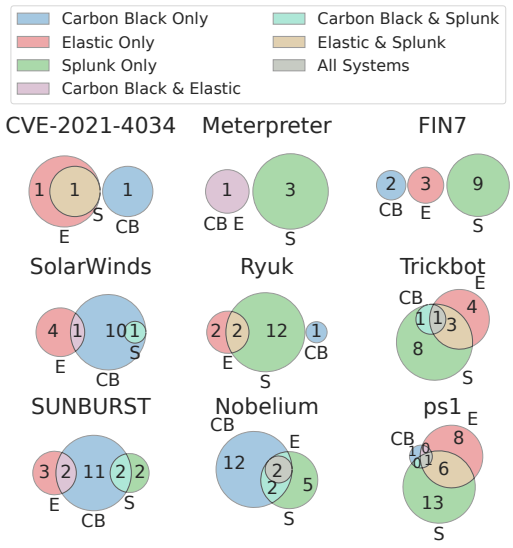

Figure 5: **Comparing ATT&CK Technique Labels for Endpoint Detection Rules Designed for the Same Malicious Entities:** These 9 malicious entities have dedicated rules in all three rule sets. ATT&CK label agreement for each entity is generally poor, suggesting there may exist multiple valid interpretations for the same threat, and it is difficult to use the ATT&CK framework to tag the threat consistently across products.

6.1 Grouping Rules across Products

To identify rules created for the same attack behavior across products, we first group rules based on their *rule metadata*. More specifically, rules are often created to counter specific malware, vulnerabilities (CVE), malicious campaigns, or threat groups (threat actors), which are usually mentioned in the rule description. We consider these specific threats to be "malicious entities". Take this description from an Elastic rule as an example: "*The malware known as SUNBURST targets the SolarWind's Orion business software for command and control. This rule detects post-exploitation command and control activity of the SUNBURST backdoor.*" Here the description references two malicious entities: the "SUNBURST backdoor" and the attack campaign against "SolarWinds" discovered in December 2020 [22].

As such, we construct a list of *malicious entities* (i.e., keywords associated with specific procedure-level threats) and perform preliminary grouping of the rules. We determine the malicious entities that a rule is associated with in the following ways. First, from the MITRE ATT&CK framework, we obtain a list of common threat groups (135), software names (718), and campaign names (14), as these often serve as reasons for rule creation. "Threat groups" in MITRE ATT&CK are typically used to refer to threat actors; MITRE also keeps track of the multiple names of the same threat actor (assigned by different security vendors). Then, we augment the list of threat groups and software names using data from the Malpedia library [25]. In total, the concatenated list contains 3,920 items. Using this item list, we search the rule metadata for potential matches. Finally, to identify additional items that are not tracked by MITRE ATT&CK or Malpedia, we further extract keywords from each rule's description field using a popular keyword extraction tool called KeyBERT [44].

Analyzing the matched entities, we note that not all of the listed items (from MITRE, Malpedia or KeyBERT) are necessarily malicious. For example, there are benign softwares such as "Reg" and "Net" that have known patterns of misuse from adversaries as well as legitimate uses by administrators and pen testers. The list also contains popular benign software such as "powershell" that is often the target of malicious attacks. To this end, we *manually* go through the results to verify that the matched items and keywords are correct, and filter to only include those that we consider "malicious entities."

Across all three rulesets, we identified and verified 191 malicious entities. These include 62 malware names and 30 threat actors (explicitly matched with the MITRE/Malpedia list), 49 CVEs, and 50 additional malicious entities from our keyword extraction (KeyBERT). This corresponds to 100 (11.5%) rules in Carbon Black, 429 (47.1%) rules in Splunk, and 84 (17.8%) rules in Elastic. Note that a rule may be associated with multiple malicious entities. For example, 34 rules are labeled with both *nobelium* and *sunburst*. This is because the Nobelium threat group used the SUNBURST backdoor in their 2020 attacks against SolarWinds [49].

Overlapping Malicious Entities. Appendix Figure 6 visualizes the overlap of dedicated rules from the three products for detecting each of the malicious entities we identify. We observe that there is little overlap between all three products with only 9 malicious entities (out of 191, 4.7%) explicitly mentioned in rules from *all three systems* and 37 malicious entities (19.3%) mentioned in at least two systems. 153 (out of 191, 80.1%) malicious entities were only mentioned in a single product. However, we *strongly emphasize* that this result does not suggest that these products are vulnerable or incomplete. It is likely that all vendors have examined threat intelligence for all of the malicious entities. In cases where dedicated rules are not present, they may have determined that their more generic detection rules were sufficient, that the creation of dedicated rules for a given threat would increase false alarms, or simply did not mention the malicious entity in the rule description. In any of these scenarios, our grouping method would have missed the fact that a malicious entity has already been accounted for. Instead, the value of this analysis is in identifying 37 malicious entities that have associated rules in more than one product to analyze how they apply the MITRE ATT&CK framework. A subset of these malicious entities include **CVE-2021-4034**, **Trickbot**, **Ryuk**, **Nobelium**, **SolarWinds**, **ps1**, **Meterpreter**, **SUNBURST**, **FIN7** – that all three products unambiguously set out to detect.

6.2 Technique Labeling Consistency Analysis

We examine the technique tags assigned by each product to the 37 common malicious entities to examine how consistently the ATT&CK framework is applied. We observe that there is generally very little agreement between the different products as to which techniques to associate with a particular threat. We report the detailed results for each malicious entity as a table in our supplementary materials [84]. Out of the 37 entities, 19 (51%) having *no agreement* in MITRE ATT&CK techniques between any pair of products. Only 1 out of the 37 entities (2.7%) have a *perfect agreement* in the techniques between the products. The result confirms the major inconsistency when applying MITRE ATT&CK to the same threat.

Among the 37 entities, 9 entities are covered by all three products' rulesets. We use Figure 5 to further visualize the level of agreement between systems for these 9 entities. We use the venn diagram to show the overlap of the technique tags for the associated rules. For instance, for the **FIN7** advanced persistent threat (APT), there is no agreement between the three products. More specifically, there are a total of 14 techniques associated with the FIN7 APT rules (9 from Splunk, 3 from Elastic, and 2 from Carbon Black), but there is no agreement about even a single technique from any of the systems under investigation. Across the 9 threats appearing in all rulesets, the three products agree on just 4 labels (2.8%) out of 141 technique annotations. When considering the 37 threats appearing in at least two rulesets, products agree on 37 (12.5%) out of 296 technique annotations. We acknowledge that threat groups and even individual malware implementations may exhibit a wide range of behavior which could contribute to differences in technique. However, within the groups of rules associated with a threat, we observe instances of products detecting the same system-level behavior where we expect to see technique agreement. In the following, we provide more detailed case studies on these rules and their ATT&CK technique labels.

CVE-2021-4034 [54]. This software vulnerability was disclosed in 2021 and describes a bug in polkit's `pkexec` utility that can be exploited to grant privilege escalation. The ATT&CK techniques that the three products associate with this vulnerability and their descriptions are shown in Table 6. This example highlights a significant issue analysts face when trying to apply MITRE ATT&CK – ambiguity and overlap between techniques. In this case, one source of disagreement is that Carbon Black uses `T1548` (Abuse Elevation Control Mechanism) while Elastic and Splunk use `T1068` (Exploitation for Privilege Escalation). The high-level descriptions provided by MITRE for these two techniques describe behavior that is difficult to distinguish, and in fact both techniques in this case appear to be a valid description of CVE-2021-4034's privilege escalation vulnerability.

Meterpreter [48]. As another case study, we consider a rule from each ruleset associated with the Meterpreter payload [48]. These rules all detect named pipe impersonation with minor differences in implementation.

EDR	Technique	Description
Carbon Blk	`T1548`	Abuse Elevation Control Mechanism
Elastic	`T1574`	Hijack Execution Flow
Elastic	`T1068`	Exploitation for Privilege Escalation
Splunk	`T1068`	Exploitation for Privilege Escalation

Table 6: **Inconsistent MITRE ATT&CK Technique Labels for CVE-2021-4034:** Techniques assigned to CVE-2021-4034 rules by different endpoint detection products.

```
event.type == "start"
and process.pe.original_file_name in ("Cmd.Exe", "
   PowerShell.EXE")
and process.args : "echo"
and process.args : ">"
and process.args : "\\\\.\\pipe\\*"
```

This rule from Elastic [21] is annotated with `T1134` (Access Token Manipulation) and tactics *Defense Evasion* and *Privilege Escalation*.

```
Processes.process_name=cmd.exe
OR Processes.original_file_name=Cmd.Exe
OR Processes.process=*%comspec%*
(Processes.process=*echo* AND Processes.process=*pipe*)
```

This rule from Splunk [78] is annotated with techniques `T1059` (Command and Scripting Interpreter) and `1543` (Create or Modify System Process) and tactics *Execution*, *Persistence*, and *Privilege Escalation*. Note that we have removed some Splunk specific syntax and only included the portion relevant to the detection here.

Carbon Black also has an overlapping rule with the same ATT&CK tags as Elastic (`T1134`). We observe that the following event, provided by Splunk [78] in the corresponding rule's description, would cause all three rules to fire:

```
cmd.exe /c echo 4sgryt3436 > \\.\pipe\5erg53
```

The technique labels for these rules are sensible, as named pipe impersonation may be used to assume the access token of the client user connected to the pipe, and here it involves misuse of the command line interpreter. Again, this example highlights ambiguity within the techniques when considering procedure-level behavior (i.e., logged system activity).

Tactic Disagreement Example. We conduct a case study of two similar rules from Elastic and Splunk that fire when DNS utility `nslookup.exe` is executed with specific command line arguments. This behavior has been associated with several actors and threats, including FIN7 and SUNBURST.

```
event.category:process
and event.type:start
and process.name:nslookup.exe
and process.args:
(-querytype=* or -qt=*
or -q=* or -type=*)
```

This rule from Elastic [21] is annotated with technique `T1071` (Application Layer Protocol) and tactic *Command and Control*.

```
Process.process_name = "nslookup.exe"
Process.process = "*-querytype=*" OR
Process.process = "*-qt=*" OR
Process.process = "*-q=*" OR
Process.process = "-type=*" OR
Process.process = "*-retry=*"
```

This rule from Splunk [78] is annotated with technique `T1048` (Exfiltration Over Alternative Protocol) and tactic *Exfiltration*. Note that we have removed some Splunk specific syntax and only included the portion relevant to the detection here.

The only difference between the rules is an additional command line argument ("*-retry=*) in the Splunk implementation of the rule. Thus there is a large overlap in the sets of system logs that would cause these rules to fire alerts. However, we again observe that there is a disagreement between products about which ATT&CK technique these rules cover. Elastic is annotated with `T1071` (Application Layer Protocol) while Splunk is annotated with `T1048` (Exfiltration Over Alternative Protocol). Both systems make reasonable decisions for technique coverage based on the high-level description. These descriptions are sufficiently general that there is no well-defined mapping between system-level behavior and the techniques associated to it.

More concerningly, we note that these two techniques fall under different tactics. Elastic tagged tactic *Command and Control* while Splunk tagged tactic *Exfiltration*. *If we take the perspective of a security analyst investigating a breach or conducting attack reconstruction, they may attribute the same system log activity to two completely different motivations depending on which product they are using.* From Elastic, we would assume this activity was aimed at gaining control of the host system, while with Splunk we would assume the goal was to steal host data.

Inconsistent Technique Labels within the same product. We also observe instances where a given product may annotate similar rules within their own product with different technique labels. For example, we consider two rules from Carbon Black associated with the NotPetya malware. Both rules detect child processes of `lsass.exe`; however, one rule specifies 4 known bad child processes, while the other looks for any child process excluding a list of known false positives. In short, the events that would cause the first rule to fire is a subset of the second rule. The first rule is only annotated with `T1547` (Boot or Logon Autostart Execution) under tactics *Persistence and Privilege Escalation*, and the second rule is annotated with `T1547` and a new `T1003` (OS Credential Dumping) under tactic *Credential Access*.

Another example for similar rules with different labels is Splunk's rules related to CVE 2021-34527 with spawned `rundll32` processes. One rule specifies that the parent process should be `spoolsv.exe`, annotated with `T1547` (Boot or Logon Autostart Execution) while the other rule is broader and does not specify a parent, annotated with `T1218` (System Binary Proxy Execution). This points to potential hierarchical relationships between different techniques within the ATT&CK Framework.

7 Disclosures

Following completion of this study, we reached out to the product vendors in an attempt to disclose our results. We successfully connected with one of the surveyed vendors' technical and marketing teams, who were "excited" to hear about this work. Contrary to our intuition that vendors may be concerned about our findings – indeed, some still might – this vendor felt that a deeper discussion of the ATT&CK framework's role in security products was a good thing. In particular, marketing staff saw our findings as an opportunity to counter-balance the emphasis on ATT&CK coverage in product reports such as Gartner's Magic Quadrant [27]. Our conversation indicated that technical and marketing staff were already aware of the potential for tension between ATT&CK coverage metrics and effective security monitoring.

We also reached out to MITRE's ATT&CK team and their Center for Threat-Informed Defense, and had an opportunity to share our results with multiple relevant MITRE staff. Although their guidelines advise against using ATT&CK techniques as a checklist to complete – warning users not to shout "Bingo" when they've covered a technique [81] – the staff we spoke to were aware of some community misconceptions surrounding the framework. The largest misconception they observed was that MITRE ATT&CK is a complete summarization of all attack techniques and behaviors–since it only includes information that has been repeatedly verified, it may not have coverage of new APT behaviors. Another guideline [81] warns against the assumption that identifying one method of performing a technique is sufficient, as attackers can have a variety of system behaviors associated with a single technique. Regarding evaluation of EDR systems, they confirmed the importance of investigating the details of low-level detection behaviors (e.g., in the Engenuity Evaluations [52]) rather than relying on coverage metrics. For this reason, one staff member argued that it may not be a problem that products inconsistently apply ATT&CK technique labels, that instead this diversity could be a positive thing.

Finally, we also discussed our results with a leading cyber risk assessment company, who provides evaluations of their customers' security posture based on hundreds of metrics. Practitioners from this company indicated that they saw the value in ATT&CK as an explanation tool for incidents rather than a predictive factor of future events. Further, they mentioned that the security community is not aligned about what "TTPs" (i.e., tactics, techniques, and procedures) are and how they happen at an endpoint. This is reflected in our findings on inconsistent ATT&CK labels across products. More concerningly, they noted a misalignment between practitioners and the industry at large—they observed that their own customers often relied on advertisements of ATT&CK coverage

to decide which security product to purchase and interpreted that coverage as a definite notion of security (i.e., 90% coverage of MITRE ATT&CK equals "90% secure"). Additionally, one staff member noted that incorporating signature-based rules with anomaly detection and confirming correct configurations and deployments was more important for security than addressing all ATT&CK techniques. This indicates that while many security practitioners are aware of how to properly use ATT&CK, others can still be influenced by its misuse in marketing materials and other communications.

8 Discussion

8.1 MITRE ATT&CK as a Security Metric

The MITRE ATT&CK framework is increasingly used as the basis for evaluating threat readiness. Owing to its systematization of threats and periodic evaluation challenges, using ATT&CK in marketing materials has become an industry-wide practice for security vendors (e.g., [15,31,70,87]). Based on this messaging, organizations regularly use ATT&CK technique coverage as a measure of the efficacy of their deployed security tools (e.g., [9,65]). The ATT&CK framework is beginning to be used for assessing cyber risk and liability, and may even be factored into calculations for some cyber insurance premiums [5]. Put another way, the MITRE ATT&CK framework is yet another (potentially problematic) security metric [23,63].

The present study highlights the many pitfalls of blindly using MITRE ATT&CK technique coverage as a security metric. We analyze three major endpoint detection products that choose not to pursue 100% coverage of ATT&CK, but instead hover at 48%-55% coverage. Even at this coverage level, many techniques are only implemented as low priority rules that vendors define as unreliable (e.g., [77]). We believe it is unlikely that vendors have been negligent in their usage of MITRE ATT&CK; instead, we interpret this as evidence that large portions of ATT&CK are not suitable to implement as endpoint detection rules. In addition, we find similar biases (with statistical significance) among vendors in their technique coverage, suggesting that vendors have similar preferences for which ATT&CK techniques to write rules.

Further undermining the notion of ATT&CK coverage as an infallible security metric is our discovery that different vendors apply the framework in different ways, through our investigation of specific threats mentioned in rules of multiple endpoint detection products. Even when all vendors were specifically attempting to detect the same malware behavior, we observed very little agreement between products as to the techniques that should be used to describe the threat. These inconsistencies at the technique level can even cause an analyst to reach the wrong conclusions about an attacker's tactical goal. While we agree with MITRE staff that, as an explanation framework, diversity (inconsistency) in ATT&CK labeling is unlikely to cause problems in the hands of experienced professionals, the problem arises when ATT&CK is instead misused as a coverage-based security metric. Given that vendors' apply the ATT&CK framework in different ways, it is unclear what "coverage" of a given technique can tell us about an organization's security posture. We identify in Section 7 that vendors have also begun to experience the negative effects of coverage as a security metric.

Our findings provide empirical evidence for anecdotal arguments against ATT&CK coverage metrics raised by practitioners. We highlight the potential implications of the lack of a strict hierarchy between techniques to tactics [69], and show how ambiguous and overlapping definitions may lead vendors to label the same behavior with conflicting technique tags. We also demonstrate that, in spite of financial incentives to inflate ATT&CK coverage as much as possible, vendors often leave large portions of the ATT&CK framework uncovered, or minimize the importance of certain techniques by assigning their rules a low priority. This apparent contradiction between marketing materials and the actual products supports the observation that many ATT&CK techniques are rarely used by adversaries or suffer from poor signal-to-noise ratios when implemented as rules [29]. In contrast with prior anecdotal narratives [69], we conduct a comprehensive qualitative analysis of the ATT&CK techniques. The findings of our systematic coding of ATT&CK techniques also support the argument that many techniques are difficult, if not impossible, to implement as detection rules [29]. Aside from attack behaviors that are not targeted at an endpoint (e.g., external infrastructure) and logically cannot be detected, endpoint detection products may suffer from false negatives related to chained attack events. Such attacks require provenance-based detection, while rules typically encode isolated behaviors. While industry blogs [29,69] discuss difficulties with usage of MITRE ATT&CK, we provide substantiated evidence to identify how they manifest in widely deployed endpoint detection products and their implications for security analysts and vendors. Aside from analyses at the tactic and technique level, we also conduct procedure-level comparisons in downstream usage of ATT&CK across products. This leads to more nuanced insight into the impact of hierarchies and ambiguities within ATT&CK. We observe that ambiguities in how ATT&CK is interpreted can lead to divergent conclusions during attack reconstruction depending on the utilized product.

Complementary to ATT&CK, MITRE has also recently introduced the D3FEND Framework to describe cybersecurity countermeasures [80], similar in objective to the NIST Cybersecurity Framework (CSF) [56]. Both D3FEND and CSF catalogue defensive cybersecurity capabilities, rather than threats. D3FEND is a newer framework, but is far more fine-grained than CSF in its description of defensive capabilities; in turn, the CSF is more process-driven and outlines how organizations can set out and achieve a target security posture. Both D3FEND and CSF are better positioned than ATT&CK

to evaluate an organization's cybersecurity preparedness. This is because they focus on defensive capabilities and procedures rather than the nature of threats. That said, neither defensive framework sets out to evaluate or prescribe specific countermeasures, so they are not an immediate remedy to how ATT&CK is used to market products. Further, there may be situations where two products offering the same capability provide different levels of security, which cannot be expressed in D3FEND or CSF.

8.2 Recommendations

In light of the above, we strongly advise vendors, practitioners, insurers, and even researchers to avoid overreliance on coverage-based evaluations of MITRE ATT&CK. The technique level of ATT&CK offers a seductive middle-ground for security evaluations — it is complex enough for coverage to seem meaningful, but ultimately still simple enough for non-experts to understand. Of course, real adversaries do not exist at the technique level of ATT&CK, but at the procedure level; techniques exist to provide generalizable descriptions of specific attack procedures observed in verified real-world incidents. Detection rules also operate at the procedure level, describing specific interactions between system entities. This fundamental disconnect between the *procedures used for detection* and the *techniques used for evaluation* may lead to misinterpretations of MITRE ATT&CK coverage analysis. We observe such misinterpretations in the advertisements discussed in Section 2. For example, Cynet's blog [18] on the 2022 MITRE ATT&CK Evaluation implies that their high coverage percentages are a good measure of system effectiveness. As a result, even if the issues of ambiguity were addressed, ATT&CK coverage statistics still lead to a false sense of security because technique coverage implies security against one of a (possibly unlimited) number of different procedure-level threats. While procedure-level coverage analysis may seem like an obvious way to address this issue, MITRE adopts a conservative approach that admits only *verified* threat groups and procedures [69]. Thus, while procedure-based analysis would bring evaluations more in line with actual product behaviors, it is also flawed because ATT&CK is not intended to be a comprehensive repository of threat intelligence.

For practitioners and rule authors, we recommend taking steps to support rule evaluation via other methods. In general, the problem of evaluating a detection rule is difficult because its performance depends on the context and environment. That said, we call for the development of feedback mechanisms that allow practitioners to share feedback about rule performance alongside the exchange of other threat intelligence. For example, many of the sources of false alarms for a given detection rule will be common across different organizations. Exchanging this information would be helpful in identifying detection rules with poor signal-to-noise ratios. Yet, we are not aware of any security product that directly collects practitioner feedback on a systematic level on whether an alert is a false alarm. The security products in our study have forums to receive this information individually, either internally or via GitHub issues. However, this collection is an ad hoc process and requires manual investigation by analysts. Further, it is difficult to synthesize feedback across products for similar rules. Open source projects like Sigma are particularly well-positioned as a repository for community-wide rule feedback. A promising direction for future work would be designing mechanisms for evaluating and providing feedback on detection rules.

Our work shows that different practitioners may assess similar rules with different ATT&CK techniques. Another future direction is a recommendation system for ATT&CK labeling that can automatically determine the appropriate set of techniques for a particular rule implementation. Depending on the context in which a rule fires, the appropriate tactic or technique could differ. In this case, it may be useful to have some dynamic assessment where the technique is assigned at the time of the alert given the surrounding system behavior. This would provide a more tailored attack contextualization to the analyst and improve the endpoint product's usability.

Just as this work is not intended as a critique of the surveyed security products, we also feel it is unfair to disproportionately blame MITRE ATT&CK for these issues. MITRE actively advises against the kinds of coverage-based analysis that has overtaken the industry, specifically emphasizing that evaluations are only a starting point, that there are no winners in their evaluations, and even that not all techniques are created equal [52]. Instead, we argue that these problems are a result of the misapplication of ATT&CK that arise as an emergent property of the security ecosystem — vendors need sales pitches, consultants need to offer actionable advice, and practitioners need ways to evaluate these claims.

Our primary recommendation to MITRE is to take a more active role in shaping how the ecosystem is (mis)using the ATT&CK framework. The available guidance for ATT&CK has focused on the point of threat intelligence creation, such as annotating malware samples with techniques. It is necessary for MITRE to more broadly disseminate how ATT&CK should and *should not* be used, especially as coverage-based analysis begins to find its way into the cyber liability and insurance industries. This is especially important given the misalignment between practitioners and the remainder of the industry, as raised during our conversation with the cyber risk assessment company (Section 7). MITRE could also consider providing more extensive guidelines about how to interpret the ATT&CK framework. We also suggest formalizing latent relationships within ATT&CK, such as hierarchies between discrete techniques. While these patterns may be realized anecdotally by individual organizations through their attack traffic, MITRE may be positioned to perform a large-scale survey or data collection to systematize this information. MITRE ATT&CK remains a fantastic knowledge base of real-world

threat behaviors, and we are confident that issues of technique ambiguity will continue to be iteratively addressed.

8.3 Limitations and Future Work

An important consideration of our findings is whether the surveyed products and rulesets are representative of the entire security ecosystem. First and foremost, we selected the rulesets in this work based on availability. The majority of top products [24, 26] are proprietary and do not make their rulesets visible to customers; instead, the existence of a rule can only be inferred when an alert is fired. We are not aware of any other public or semi-public commercial rulesets, but hope to expand our results in future work as more become available. Whether or not they are representative of other products, Splunk, Elastic, and Carbon Black combine to account for a huge proportion of the ecosystem, having been deployed on millions of machines in thousands of organizations from various regions and industries [61,76,85]. Thus, it is not necessary for our findings to be universally applicable to have important implications for real-world security. Further, our analysis of the crowdsourced Sigma ruleset, which is comprised of rules submitted by hundreds of contributors and has downstream use by industry partners such as IBM [72], appear to confirm the general trends observed in our main analysis.

Our work is focused on endpoint detection; it is a critical component of enterprise security, but may not represent other systems such as network threat detection. While conducting this study, we attempted to collect and analyze detection rules from Network Detection & Response (NDR) rulesets. However, of the public rulesets we obtained, we found that ATT&CK technique annotations were far less common than for endpoint detection rules. It is not clear whether this observation is generally true of network detection rules or is simply an artifact of our limited visibility into this ecosystem. Future work is needed to extend our analysis to network detection.

Another threat to validity is the possibility that our qualitative analysis (§5) may have been biased by the coders' background and expertise, which is inherent to this type of analysis. We mitigate this concern by using three coders and by taking a conservative approach; specifically for the "ineffective detection method" coding, we only classify techniques as such when explicitly mentioned in MITRE's text, or if its detection would *obviously* result in high false positives.

Finally, our method of grouping rules by malicious entity (§6) is *necessarily* incomplete. It may be the case that some rules were incorrectly excluded from a group because their description did not explicitly mention the threat entity. It is certainly the case that many rules were excluded because they were designed to capture a more general class of malicious behavior. We made this trade-off to ensure the soundness of the analysis – by being conservative in our assignment of rules to a given threat entity, we ensured that the rule was unambiguously intended to detect that entity. This lead to the most favorable conditions for consistency as different rule authors are assigning MITRE ATT&CK annotations; yet, in spite of this, we still discovered widespread inconsistency amongst this subset of rules. Future work may explore a more generalizable approach to comparing rules at the procedure level to expand this analysis.

9 Related Work

MITRE ATT&CK. Researchers have proposed to map vulnerabilities to MITRE ATT&CK using machine learning techniques [40, 47, 68, 88], and use ATT&CK to characterize security risks and facilitate threat modeling [2, 67]. Prior work has also analyzed malware [60] and threat intelligence reports [64] to identify common ATT&CK techniques used in practice. A recent work [66] systematizes ATT&CK research and highlight its use cases, application domains, and related frameworks. Our analysis has important implications for work that performs ATT&CK-based threat modeling and risk assessment, especially those that assume the framework is an exhaustive and uniformly-likely enumeration of possible attack behaviors. Based on our findings and how the knowledge base is constructed, we urge authors to avoid conflating ATT&CK tactics, techniques, and procedures as "TTPs."

EDR Systems. MITRE ATT&CK has been prominently figured into research that integrates data provenance analysis with traditional EDR [10, 19, 32, 33, 41, 51]. Systems like HOLMES [51] and RapSheet [32] specifically assume that the underlying EDR will generate alerts associated with every ATT&CK tactic; however, our analysis indicates that this is highly unlikely to occur in practice. Qualitative studies [3, 39] have identified usability issues caused by false positives. However, as shown in our analysis, disabling rules with high false alarms (i.e., low-confidence rules) further reduces EDR's coverage of ATT&CK. Our work is orthogonal to research on *false negatives* in EDR products [38, 58]. For example, Karantzas et al. [38] evaluate eleven EDR products against four attack scenarios and find that all the EDRs fail to detect at least one attack. Complementary to our work, Shen et al. [71] investigate the implications of the MITRE Engenuity evaluations on real world EDR performance.

Threat Detection and Intelligence. Thematically similar to our study, Bailey et al. [8] compared Internet malware classifications behaviors of various signature-based antivirus products. Prior work has also compared various sources in the threat intelligence community [11, 30, 43], evaluating metrics like latency [43]. and originality [30]. Several papers [4, 6, 7, 28, 82, 83] have performed comparative analysis of rule-based network intrusion detection systems (NIDS), with a focus on their performance against popular attacks [4, 82], and the potential overlap [7, 83] and evolution [83] of NIDS rulesets. However, different from our analysis, these studies did not focus on the application of MITRE ATT&CK on the rulesets.

10 Conclusion

We present a comprehensive analysis of how the MITRE ATT&CK framework is used across four widely deployed endpoint detection products. We find that *ATT&CK coverage* is inflated by the presence of low-risk rules and different vendors classifying rules describing the same system-level behavior with inconsistent ATT&CK techniques and tactics. The results indicate that *ATT&CK coverage* may not be a suitable security metric for evaluating endpoint detection products. We conclude by providing recommendations for endpoint detection vendors and MITRE to improve the usage of ATT&CK.

Acknowledgments

We thank the reviewers for their helpful suggestions. This work was supported in part by NSF grants 2055127, 2229876, and 1955719, as well as the IBM-ILLINOIS Discovery Accelerator Institute and VMware University Research Fund.

References

[1] 0xAnalyst. CB-Threat-Hunting. https://github.com/0xAnalyst/CB-Threat-Hunting, Last accessed June 2023.

[2] M. Ahmed, S. Panda, C. Xenakis, and E. Panaousis. MITRE ATT&CK-Driven Cyber Risk Assessment. In *International Conference on Availability, Reliability and Security*, 2022.

[3] B. A. Alahmadi, L. Axon, and I. Martinovic. 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. In *USENIX Security Symposium*, 2022.

[4] E. Albin and N. C. Rowe. A realistic experimental comparison of the Suricata and Snort intrusion-detection systems. In *International Conference on Advanced Information Networking and Applications Workshops*, 2012.

[5] A. Antoni and P. Dyson. Cyber Risk Quantification based on the MITRE ATT&CK Framework. https://www.kovrr.com/blog-post/cyber-risk-quantification-based-on-the-mitre-attck-framework, March 2023.

[6] H. Asad and I. Gashi. Diversity in Open Source Intrusion Detection Systems. In *Computer Safety, Reliability, and Security*, 2018.

[7] H. Asad and I. Gashi. Dynamical analysis of diversity in rule-based open source network intrusion detection systems. *Empirical Software Engineering*, 27:1–30, 2022.

[8] M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, and J. Nazario. Automated classification and analysis of internet malware. In *Recent Advances in Intrusion Detection*, 2007.

[9] J. Basra and T. Kaushik. MITRE ATT&CK as a Framework for Cloud Threat Investigation. https://cltc.berkeley.edu/publication/mitre-attck/, Oct 2020.

[10] B. Bhattarai and H. Huang. Steinerlog: prize collecting the audit logs for threat hunting on enterprise network. In *ACM Asia Conference on Computer and Communications Security*, 2022.

[11] X. Bouwman, H. Griffioen, J. Egbers, C. Doerr, B. Klievink, and M. Van Eeten. A different cup of TI? The added value of commercial threat intelligence. In *USENIX Security Symposium*, 2020.

[12] M. Buchanan. Rapid7 2023 MITRE Engenuity ATT&CK® Evaluations. https://www.rapid7.com/blog/post/2023/09/20/rapid7-delivers-visibility-across-all-19-steps-of-attack-in-2023-mitre-engenuity/, 2023.

[13] Business Wire. Stellar Cyber Launches MITRE ATT&CK Coverage Analyzer for Partners and Customers. https://www.businesswire.com/news/home/20240423051638/en/, 2024.

[14] Carbon Black. Carbon Black Delivers MITRE ATT&CK™ Coverage with Zero Delayed Detections & Zero Tainted Detections. https://news.vmware.com/releases/carbon-black-delivers-mitre-attck-coverage-with-zero-delayed-detections-zero-tainted-detections, 2018.

[15] Check Point Software Technologies Ltd. MITRE ATT&CK Coverage. https://www.checkpoint.com/solutions/mitre-attack/coverage/, 2024.

[16] R. Chheda. Our Take: SentinelOne's 2022 MITRE ATT&CK Evaluation Results. https://www.sentinelone.com/blog/our-take-sentinelones-2022-mitre-attck-evaluation-results/, 2022.

[17] Cybersecurity and Infrastructure Security Agency. Best Practice for MITRE ATT&CK Mapping. https://www.cisa.gov/sites/default/files/publications/Best%20Practices%20for%20MITRE%20ATTCK%20Mapping.pdf, June 2021.

[18] Cynet. Learn how to interpret the 2022 MITRE ATT&CK Evaluation results. https://www.cynet.com/blog/learn-how-to-interpret-the-2022-mitre-attck-evaluation-results/, 2022.

[19] F. Dong, S. Li, P. Jiang, D. Li, H. Wang, L. Huang, X. Xiao, J. Chen, X. Luo, Y. Guo, et al. Are we there yet? An Industrial Viewpoint on Provenance-based Endpoint Detection and Response Tools. In *ACM Conference on Computer and Communications Security*, 2023.

[20] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank Aggregation Methods for the Web. In *World Wide Web Conference*, 2001.

[21] Elasticsearch B.V. Detection Rules. https://github.com/elastic/detection-rules, 2022.

[22] FireEye. Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims With SUNBURST Backdoor. https://www.mandiant.com/resources/blog/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor, 2020.

[23] D. Flater. Bad security metrics part 1: Problems. *IT Professional*, 20(1):64–68, feb 2018.

[24] Forrester. The Forrester Wave™: Endpoint Security, Q4 2023. https://www.forrester.com/report/the-forrester-wave-tm-endpoint-security-q4-2023/RES178486, 2023.

[25] Fraunhofer FKIE. Malpedia. https://malpedia.caad.fkie.fraunhofer.de/, 2023.

[26] Gartner. Endpoint Detection and Response (EDR) Solutions Reviews 2023. https://www.gartner.com/reviews/market/endpoint-detection-and-response-solutions, 2023.

[27] Gartner. Gartner Magic Quadrant for Endpoint Protection Platforms. https://www.gartner.com/en/documents/4001307, 2023.

[28] H. Gascon, A. Orfila, and J. Blasco. Analysis of update delays in signature-based network intrusion detection systems. *Computers and Security*, 30(8):613–624, 2011.

[29] C. Gerritz. Why you're going about MITRE ATT&CK coverage all wrong. https://securityboulevard.com/2021/03/why-youre-going-about-mitre-attck-coverage-all-wrong/, March 2021.

[30] H. Griffioen, T. Booij, and C. Doerr. Quality evaluation of cyber threat intelligence feeds. In *Applied Cryptography and Network Security*, 2020.

[31] C. Hankins. Assessing and expanding MITRE ATT&CK coverage in Splunk Enterprise Security. https://lantern.splunk.com/?title=Security/UCE/Guided_Insights/Cyber_frameworks/Assessing_and_expanding_MITRE_ATT%26CK_coverage_in_Splunk_Enterprise_Security, 2024.

[32] W. U. Hassan, A. Bates, and D. Marino. Tactical Provenance Analysis for Endpoint Detection and Response Systems. In *IEEE Symposium on Security and Privacy*, 2020.

[33] W. U. Hassan, S. Guo, D. Li, Z. Chen, K. Jee, Z. Li, and A. Bates. Nodoze: Combatting threat alert fatigue with automated provenance triage. In *Network and Distributed Systems Security Symposium*, 2019.

[34] P. Havens. 2022 MITRE Engenuity ATT&CK Evaluations Results. https://www.paloaltonetworks.com/blog/2022/03/mitre-engenuity-evaluations-round-4-results/, 2022.

[35] E. M. Hutchins, M. J. Cloppert, and R. M. Amin. Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains, 2011.

[36] C. Iordache. MITRE ATT&CK Evaluations 2023: Deciphering the Results. https://www.bitdefender.com/blog/businessinsights/mitre-attck-evaluations-2023/, 2023.

[37] K. Jagtap. SafeBreach Enhances ATT&CK Coverage with Industry Scenarios Focused on Top-16 MITRE TTPs. https://www.safebreach.com/blog/top-16-mitre-attack-ttps/, 2022.

[38] G. Karantzas and C. Patsakis. An Empirical Assessment of Endpoint Detection and Response Systems against Advanced Persistent Threats Attack Vectors. *Journal of Cybersecurity and Privacy*, 1(3):387–421, 2021.

[39] F. B. Kokulu, A. Soneji, T. Bao, Y. Shoshitaishvili, Z. Zhao, A. Doupé, and G.-J. Ahn. Matched and Mismatched SOCs: A Qualitative Study on Security Operations Center Issues. In *ACM Conference on Computer and Communications Security*, 2019.

[40] A. Kuppa, L. Aouad, and N.-A. Le-Khac. Linking CVE's to MITRE ATT&CK Techniques. In *International Conference on Availability, Reliability and Security*, 2021.

[41] K. Kurniawan, A. Ekelhart, E. Kiesling, G. Quirchmayr, and A. M. Tjoa. KRYSTAL: Knowledge graph-based framework for tactical attack discovery in audit data. *Computers and Security*, 121:102828, 2022.

[42] M. Levinson. MITRE ATT&CK Framework: 5 Questions to Ask NDR Providers about their Coverage. https://www.extrahop.com/blog/extrahop-revealx-mitre-att-ck-coverage-2024, 2024.

[43] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage. Reading the tea leaves: A comparative analysis of threat intelligence. In *USENIX Security Symposium*, 2019.

[44] MaartenGr. KeyBERT. https://github.com/MaartenGr/KeyBERT, 2023.

[45] J. Mancini. MITRE ATT&CK Coverage: Vectra AI provides over 90%. https://www.vectra.ai/blog/mitre-attack-coverage-vectra-ai-provides-over-90, 2022.

[46] N. McDonald, S. Schoenebeck, and A. Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[47] O. Mendsaikhan, H. Hasegawa, Y. Yamaguchi, and H. Shimada. Automatic mapping of vulnerability information to adversary techniques. In *International Conference on Emerging Security Information, Systems and Technologies*, 2020.

[48] Metasploit. Meterpreter. https://docs.metasploit.com/docs/using-metasploit/advanced/meterpreter/meterpreter.html, 2023.

[49] Microsoft Cyber Defense Operations Center. Deep dive into the Solorigate second-stage activation: From SUNBURST to TEARDROP and Raindrop. https://www.microsoft.com/en-us/security/blog/2021/01/20/deep-dive-into-the-solorigate-second-stage-activation-from-sunburst-to-teardrop-and-raindrop/, 2021.

[50] Microsoft Security Team. MITRE Engenuity ATT&CK® Evaluation proves Microsoft Defender for Endpoint stops advanced attacks across platforms. https://www.microsoft.com/en-us/security/blog/2021/04/21/mitre-engenuity-attck-evaluation-proves-microsoft-defender-for-endpoint-stops-advanced-attacks-across-platforms/, 2021.

[51] S. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. Venkatakrishnan. HOLMES: Real-time APT Detection through Correlation of Suspicious Information Flows. In *IEEE Symposium on Security and Privacy*, 2019.

[52] MITRE Engenuity. ATT&CK Evaluations: Using Evaluations. https://mitre-engenuity.org/cybersecurity/attack-evaluations/using-attack-evaluations/, 2023.

[53] S. Murphy. Decoding Turla: Trend Micro's MITRE Performance. https://www.trendmicro.com/en_ca/research/23/i/mitre-attack-solution-tested.html, 2023.

[54] National Institute of Standards and Technology. NVD - CVE-2021-4034. https://nvd.nist.gov/vuln/detail/cve-2021-4034, 2023.

[55] National Institute of Standards and Technology. NVD - Vulnerability Metrics. https://nvd.nist.gov/vuln-metrics/cvss#, 2023.

[56] National Institute of Standards and Technology. The NIST Cybersecurity Framework (CSF) 2.0. https://www.nist.gov/cyberframework, 2024.

[57] P. Neray. Eliminate coverage gaps with automation and MITRE ATT&CK. https://cardinalops.com/whitepapers/eliminate-coverage-gaps-with-automation-and-mitre-attck/, 2023.

[58] A. Niakanlahiji, J. Wei, M. R. Alam, Q. Wang, and B.-T. Chu. Shadowmove: A stealthy lateral movement strategy. In *USENIX Security Symposium*, 2020.

[59] M. Nichols and M. Paquette. Elastic continues to gain momentum in SIEM market. https://www.elastic.co/blog/elastic-continues-to-gain-momentum-in-siem-market, 2022.

[60] K. Oosthoek and C. Doerr. SoK: ATT&CK Techniques and Trends in Windows Malware. In *Security and Privacy in Communication Networks*, 2019.

[61] M. Paquette. Elastic named a Major Player in the IDC MarketScape: Worldwide SIEM 2022 Vendor Assessment. https://www.elastic.co/blog/elastic-named-major-player-idc-marketscape-worldwide-siem-2022, 2019.

[62] Pentera. Automatically test your entire IT infrastructure - internal and external attack surfaces. https://x.com/penterasec/status/1760573300667740280, 2024.

[63] S. Pfleeger and R. Cunningham. Why measuring security is hard. *IEEE Security & Privacy*, 8(4):46–54, 2010.

[64] M. R. Rahman, S. K. Basak, R. M. Hezaveh, and L. Williams. Attackers reveal their arsenal: An investigation of adversarial techniques in cti reports. *arXiv preprint arXiv:2401.01865*, 2024.

[65] Reality Net System Solutions. attack-coverage: An excel-centric approach for managing the MITRE ATT&CK tactics and techniques. https://github.com/RealityNet/attack-coverage, Nov 2020.

[66] S. Roy, E. Panaousis, C. Noakes, A. Laszka, S. Panda, and G. Loukas. SoK: The MITRE ATT&CK Framework in Research and Practice. arXiv, 2023.

[67] L. Sadlek, P. Čeleda, and D. Tovarňák. Current Challenges of Cyber Threat and Vulnerability Identification Using Public Enumerations. In *International Conference on Availability, Reliability and Security*, 2022.

[68] M. S. I. Sajid, J. Wei, B. Abdeen, E. Al-Shaer, M. M. Islam, W. Diong, and L. Khan. SODA: A System for Cyber Deception Orchestration and Automation. In *Annual Computer Security Applications Conference*, 2021.

[69] M. Schneider. MITRE ATT&CK: Flaws of the Standardization. https://www.scip.ch/en/?labs.20210204, 2021.

[70] M. Sentonas. CrowdStrike Achieves 99% Detection Coverage in First-Ever MITRE ATT&CK Evaluations for Security Service Providers. https://www.crowdstrike.com/blog/crowdstrike-achieves-99-percent-detection-coverage-in-mitre-attack-evaluations-for-security-service-providers/, November 2022.

[71] X. Shen, Z. Li, G. Burleigh, L. Wang, and Y. Chen. Decoding the mitre engenuity att&ck enterprise evaluation: An analysis of edr performance in real-world environments. *arXiv preprint arXiv:2401.15878*, 2024.

[72] Sigma. Main Sigma Rule Repository. https://github.com/SigmaHQ/sigma, 2022.

[73] SnapAttack. SnapAttack. https://www.snapattack.com/, 2024.

[74] Sophos. Results from the 2023 MITRE Engenuity ATT&CK Evaluations (Round 5: Turla). https://news.sophos.com/en-us/2023/09/20/results-from-the-2023-mitre-engenuity-attck-evaluations-round-5-turla/, 2023.

[75] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, XV:72–101, 1904.

[76] Splunk. The State of Security. https://www.splunk.com/en_us/form/state-of-security.html, 2023.

[77] Splunk. How are risk score calculated for RBA. https://github.com/splunk/security_content/wiki/How-are-risk-score-calculated-for-RBA, 2024.

[78] Splunk Inc. Splunk Security Content. https://github.com/splunk/security_content, 2022.

[79] The MITRE Corporation. MITRE ATT&CK®. https://attack.mitre.org, 2022.

[80] The MITRE Corporation. MITRE D3FEND™. https://d3fend.mitre.org/, 2023.

[81] The MITRE Corporation. Get Started. https://attack.mitre.org/resources/#not-use-attack, 2024.

[82] K. Thongkanchorn, S. Ngamsuriyaroj, and V. Visoottiviseth. Evaluation studies of three intrusion detection systems under various attacks and rule sets. In *International Conference of IEEE Region 10*, 2013.

[83] M. Vermeer, M. van Eeten, and C. Gañán. Ruling the Rules: Quantifying the Evolution of Rulesets, Alerts and Incidents in Network Intrusion Detection. In *ACM Asia Conference on Computer and Communications Security*, 2022.

[84] A. Virkud, M. A. Inam, A. Riddle, J. Liu, G. Wang, and A. Bates. Supplementary Materials. https://github.com/avirkud/endpoint-detection-mitreattack, 2024.

[85] VMware. Carbon Black Announces Second Quarter 2019 Financial Results. https://news.vmware.com/releases/carbon-black-announces-second-quarter-2019-financial-results, 2019.

[86] VMware, Inc. VMware Carbon Black EDR. https://www.vmware.com/products/endpoint-detection-and-response.html, 2022.

[87] VMWare Security Blog. MITRE ATT&CK Evaluation Demonstrates the Power of the VMware Carbon Black Cloud. https://blogs.vmware.com/security/2020/04/mitre-attck-evaluation-demonstrates-the-power-of-the-vmware-carbon-black-cloud.html, April 2020.

[88] F. Özdemir Sönmez, C. Hankin, and P. Malacaria. Attack Dynamics: An Automatic Attack Graph Generation Framework Based on System Topology, CAPEC, CWE, and CVE Databases. *Computers and Security*, 123:102938, 2022.

TID	Technique (*Tactics*)
T1021	Remote Services (*Lateral Movement*)
T1047	Windows Management Instrumentation (*Execution*)
T1053	Scheduled Task/Job (*Execution*, *Persistence*, *Privilege Escalat.*)
T1059	Cmd. & Scripting Interpreter (*Execution*)
T1218	Signed Binary Proxy Execution (*Defense Evasion*)
T1543	Create or Modify Sys Process (*Persistence*, *Privilege Escalat.*)

Table 7: **Multiple Technique Annotations of a Single Rule:** One Splunk rule with six technique annotations.

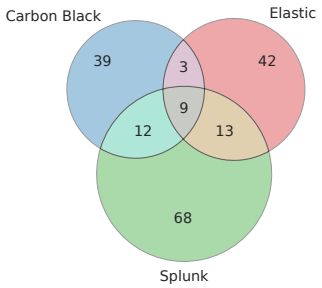

Figure 6: **Malicious Entity Coverage by Vendors:** Comparison of vendors including rules that *specifically* detect each of the malicious entities. Only 9 entities appear in all 3 rule sets. Note that limited overlap does not necessarily indicate insecurity, since EDRs may have other generic rules (that are not specifically designed for the given malicious entities) to detect these threats.

A Sigma

In this section, we replicate RQ1 and characterize how the Sigma threat detection ruleset [72] uses MITRE ATT&CK. As a crowdsourced ruleset, the quality control process for individual rules may be less uniform. This may introduce higher variation in how MITRE ATT&CK is used within the ruleset and thus we report the results separately.

First, we see from Table 1 that Sigma's coverage of ATT&CK techniques (79%) is over 20% greater than the coverage for any of the three commercial rulesets (48%-55%). Despite the higher coverage, we note that the Sigma ruleset follows similar trends in terms of technique coverage under each tactic, as shown in Figure 7. While it follows that there are fewer techniques with 0 implemented rules compared to the other EDRs, we observe in Figure 8 that Sigma also has a similar distribution of rules per technique. That is, the majority of techniques are covered by a handful of rules each.

Further, we find that the inclusion of Sigma does not change the top ATT&CK techniques across the four engines as seen in Figure 9. We replicate the Spearman coefficient calculations between the ranked list of techniques implemented by Sigma and the other three rulesets to identify if Sigma chooses to prioritize similar techniques. The Spearman coefficients are positive—0.701, 0.762, and 0.793 for Splunk, Elastic, and

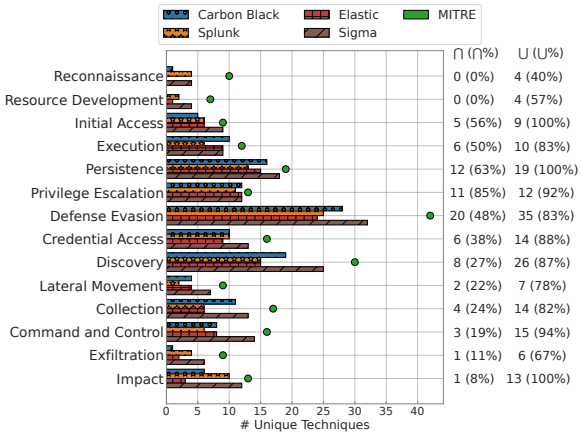

Figure 7: **Technique Coverage under Each Tactic:** The plot replicates Figure 1 with the inclusion of the Sigma ruleset.

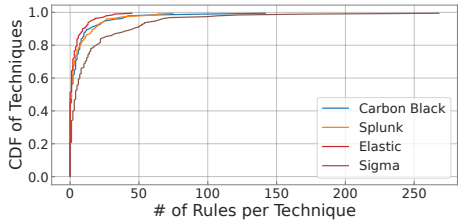

Figure 8: **Rules Per Technique:** Cumulative distribution of the number of rules per technique for each EDR system.

Carbon Black respectively—with a p-value < 0.001 for each corresponding t-test. This confirms that Sigma's similarity to all of the commercial rulesets is statistically significant in terms of which techniques each chooses to cover.

Each Sigma rule is annotated with a qualitative *level* indicating how critical the fired alert would be (i.e., how quickly a security analyst should respond), analogous to the risk and severity metrics provided by the commercial rulesets. There are five criticality levels (informational, low, medium, high, critical) and the majority of Sigma rules are assigned to either medium (34.7%) or high (48.8%) levels. We map the five levels to numeric values based on the CVSS scale [55] to investigate the level distribution for the top techniques (see Supplementary Materials [84] for figure) and find that it is fairly consistent across techniques. Similar to the commercial rulesets, we observe a drop in ATT&CK technique coverage when filtering out lower criticality level rules. If we consider only rules with at least a medium criticality level, the MITRE ATT&CK coverage slightly drops to 74%. If we consider only rules with at least a high criticality level, the MITRE ATT&CK coverage drops to 62%. Sigma also has a similar proportion of rules annotated with multiple ATT&CK techniques, as shown in Figure 10.

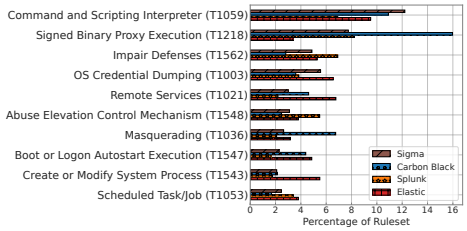

Figure 9: **Top ATT&CK Techniques:** Top 10 techniques ranked based on the sum of associated rules from the four engines. The x-axis shows the percentage of the rules for a given technique in each engine (e.g., 16% of the Carbon Black rules are tagged with `T1218`).

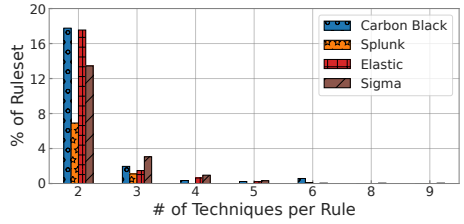

Figure 10: **Multiple Techniques Per Rule:** The distribution of the number of techniques per rule. Rules with a single technique label are omitted from the plot.

B Examples of Rules with High False Positives

```
process.name : "cmd.exe"
and event.type == "start"
and not cidrmatch(destination.ip,
"10.0.0.0/8", "127.0.0.0/8", ...)
```

This Elastic rule identifies `cmd.exe` making a network connection. The metadata stipulates that administrators may trigger this rule frequently for benign and regular tasks, causing false positives. The rule is annotated with techniques `T1059` (Command and Scripting Interpreter) linked to tactic `TA0002` (Execution), and `T1105` (Ingress Tool Transfer) linked to tactic `TA0011` (Command and Control). The severity assessment is 2.1 / 10 and the risk score is 21 / 100.

```
Processes.process_name=xclip
AND Processes.process IN ("*-o *", "*-sel *",
"*-selection *", "*clip *","*clipboard*")
```

This Splunk rule identifies the Linux tool xclip being used to copy data from the clipboard. This is commonly used by administrators and end users on Linux machines. This rule is annotated with technique `T1115` (Clipboard Data) linked to tactic `TA0009` (Collection). The confidence score is 40 / 100 and the risk score is 16 / 100.

```
Processes.process_name=curl
OR Processes.process_name=wget
```

This Splunk rule identifies usage of command-line tools curl and wget. The rule used in isolation will lead to false positives, as the behavior is likely to occur frequently in normal circumstances. This rule is annotated with technique `T1105` (Ingress Tool Transfer) linked to tactic `TA0011` (Command and Control). The confidence score is 10 / 100 and the risk score is 1 / 100.