Decoding the MITRE Engenuity ATT&CK Enterprise Evaluation: An Analysis of EDR Performance in Real-World Environments

Xiangmin Shen
Northwestern University
Evanston, Illinois, USA
xiangminshen2019@u.northwestern.edu

Zhenyuan Li
Zhejiang University
Hangzhou, Zhejiang, China
lizhenyuan@zju.edu.cn

Graham Burleigh
Northwestern University
Evanston, Illinois, USA
grahamburleigh2022@u.northwestern.edu

Lingzhi Wang
Northwestern University
Evanston, Illinois, USA
lingzhiwang2025@u.northwestern.edu

Yan Chen
Northwestern University
Evanston, Illinois, USA
ychen@northwestern.edu

ABSTRACT

Endpoint detection and response (EDR) systems have emerged as a critical component of enterprise security solutions, effectively combating endpoint threats like APT attacks with extended lifecycles. In light of the growing significance of endpoint detection and response (EDR) systems, many cybersecurity providers have developed their own proprietary EDR solutions. It's crucial for users to assess the capabilities of these detection engines to make informed decisions about which products to choose. This is especially urgent given the market's size, which is expected to reach around 3.7 billion dollars by 2023 and is still expanding. MITRE is a leading organization in cyber threat analysis. In 2018, MITRE started to conduct annual APT emulations that cover major EDR vendors worldwide. Indicators include telemetry, detection and blocking capability, etc. Nevertheless, the evaluation results published by MITRE don't contain any further interpretations or suggestions.

In this paper, we thoroughly analyzed MITRE evaluation results to gain further insights into real-world EDR systems under test. Specifically, we designed a whole-graph analysis method, which utilizes additional control flow and data flow information to measure the performance of EDR systems. Besides, we analyze MITRE evaluation's results over multiple years from various aspects, including detection coverage, detection confidence, detection modifier, data source, compatibility, etc. Through the above studies, we have compiled a thorough summary of our findings and gained valuable insights from the evaluation results. We believe these summaries and insights can assist researchers, practitioners, and vendors in better understanding the strengths and limitations of mainstream EDR products.

CCS CONCEPTS

• **Security and privacy** → **Systems security**; • **General and reference** → **Evaluation**.

ASIA CCS '24, July 1–5, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0482-6/24/07.
https://doi.org/10.1145/3634737.3645012

KEYWORDS

EDR System Evaluation, APT Emulation, Measurement Study

ACM Reference Format:
Xiangmin Shen, Zhenyuan Li, Graham Burleigh, Lingzhi Wang, and Yan Chen. 2024. Decoding the MITRE Engenuity ATT&CK Enterprise Evaluation: An Analysis of EDR Performance in Real-World Environments. In *ACM Asia Conference on Computer and Communications Security (ASIA CCS '24), July 1–5, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3634737.3645012

1 INTRODUCTION

The digital revolution dramatically changed human life and brings new risks into daily life. Driven by profit, attackers in cyberspace have organized increasingly sophisticated attacks that affect many organizations and large corporations such as Siemens, Target, and Equifax. These attacks resulted in millions of consumers' data being leaked [7, 9, 11] and other losses. Traditional network-based prevention and detection approaches can barely deal with these advanced attacks. Therefore, endpoint-based detection and response solutions (EDR) and extended solutions (XDR) receive extensive attention in academia and industry. The market capitalization for the top 37 EDR companies has reached over 320 billion U.S. dollars [12] by 2022, the market size of 3.7 billion U.S. dollars [1]. Meanwhile, both market capitalization and size are expected to grow fast.

As the number of homogeneous EDR products increases, it becomes increasingly difficult for users to choose the appropriate product. Fair third-party evaluations with detailed interpretations of results are therefore necessary. The challenges of conducting such evaluations are three-fold. Firstly, the evaluation methodology should be general enough to allow broad participation. The significance of the evaluation will be affected if its methodology only applies to a small set of security solutions. Secondly, the evaluation should perform realistic and various attack emulations. The attack emulations in the laboratory setting are simple and have little variety. Sometimes, the attack information is even known before the evaluation, making it possible for the defense team to make ad-hoc configuration adjustments. Such attack emulations can not reveal the actual performance of endpoint security solutions against real-world threats. Finally, the evaluation results should be reported with comprehensive interpretation. Without appropriate metrics and objective interpretation, the evaluation results are hard to understand, possibly leading to biased interpretation.

Although much effort has been made toward establishing a norm for security solutions, there is still no substantial benchmark in the security field. Several for-profit companies and organizations present their own evaluation results [6, 19]. However, their methodologies are not transparent and could be biased for commercial reasons. In academia, several recent benchmark works [3, 15–17, 45] focus on generating new or improving existing datasets. Although they are crucial initial steps, equally-important interpretations of evaluation results are still missing. Several other benchmarking works [22, 30] attempt to expand the explainability of evaluation results. But their methodologies are specific to the target systems or platforms, making their evaluation methodologies and results not transferable.

To standardize security evaluations, the MITRE Corporation has been conducting annual APT emulations to evaluate various security solutions based on its ATT&CK framework [33] since 2018. In each round of evaluation, MITRE will select one or more real-world APT groups and reconstruct their typical attack chains in controlled environments. The EDR products will be deployed in these environments in advance. And the results will be collected and published by MITRE to summarize their performance. The results list the attack steps characterized by MITRE ATT&CK techniques in attack emulations. For each EDR system, MITRE publishes its detection and protection performance on attack steps. The detection performance is described with MITRE-defined detection categories, indicating the amount of contextual information the EDR system provides with the alarm. The protection performance is illustrated by the step at which the attack is blocked.

While the datasets from MITRE's evaluation are valuable, the presentation of evaluation results has some apparent defects, preventing security practitioners from benefiting directly. These problems include missing whole-graph analysis, lacking comprehensive interpretation, and inconsistent evaluation framework. Concretely, MITRE only focuses on single-step detection results. However, the attacks that EDR systems fight against are sophisticated and involve multiple steps. EDR systems must consider the entire kill chain to provide satisfying detection and response services. Therefore, the whole-graph analysis capability is crucial in evaluating EDR systems. Apart from conducting attack emulations and collecting results, interpreting the results is equally or not more critical to EDR system evaluation. However, MITRE makes minimal effort in interpreting the results. Lacking comprehensive interpretation prevents users from getting direct insights from the results and could lead to biased interpretations for vendors and customers. MITRE has only conducted four evaluations so far. Understandably, its methodology has evolved, leading to several inconsistencies in the published results. Inconsistencies like these can be burdensome for users, forcing users to investigate the difference in every year's methodology.

To address these problems, we propose analysis methodologies on the MITRE evaluation dataset to perform fine-grained whole-graph analysis and holistic assessments of EDR systems' capabilities. Then, we apply our methodology to analyze MITRE evaluation datasets. We investigate all attack scenarios and construct causal relationship attack graphs to present causal relationships between attack steps. We evaluate EDR systems' attack reconstruction capability by conducting the connectivity analysis, examining whether the EDR system can reconstruct the complete attack kill chain. We also assess EDR systems' response capability via the effectiveness analysis. In the effectiveness analysis, we use protection performance as an indicator. Specifically, we examine at which step each EDR system responds to the attack and determine if the EDR system is effectively protecting the host. Moreover, we discuss the evaluation results from several practical perspectives to measure the detection and protection performance of individual techniques and EDR systems, including detection coverage, detection confidence, detection quality, data source, and compatibility. We also investigate the trend of performance change from these perspectives.

In summary, this paper makes the following contributions:

- We design and implement new analysis methods to systemically interpret MITRE ATT&CK evaluation's results, with evaluation dimensions including whole-graph analysis that explores the correlation capability of EDR systems and additional metrics to capture aspects of the evaluation results not covered by MITRE.
- We reconstruct several attack scenarios used in MITRE evaluation and apply whole-graph analysis to examine EDR systems' attack reconstruction and behavior correlation capabilities, which reveal whether an EDR system can effectively detect and respond to attacks.
- We propose a new evaluation metric and identify and highlight flaws in EDR systems. We also pinpoint a list of findings to shed light on areas that require improvement and offer suggestions to enhance the performance of EDR systems.

2 BACKGROUND

2.1 MITRE ATT&CK Evaluation

MITRE ATT&CK Evaluation is an APT emulation conducted yearly by MITRE Corporation, started in 2018. Its participants include most leading security companies, such as Palo Alto Networks, Fortinet, and CrowdStrike. Each evaluation emulates attacks from well-known APT groups like APT3, APT29, and FIN7. Contrary to other attacks like malware and phishing, APT attacks are more complicated, involving multiple stages aiming for specific tasks. Together, those stages form a kill chain to achieve the final goals, such as stealing sensitive information or destroying valuable properties. The MITRE Corporation has established a set of Tactics, Techniques, and Procedures (TTPs) [34] to outline each stage of the emulation process, which serve as a foundation for organizing steps in a kill chain. Tactics divide attack steps into 14 general stages, while techniques further distinguish attack steps according to the specific approach. In some cases, each technique can have associated sub-techniques, with additional details necessary to identify them accurately. Attacks performed in evaluations are illustrated step-wise, with individual steps associated with the techniques described above. Additionally, the information provided for each step in detection tests includes detection categories and modifiers, if applicable. The detection categories include the following types.

(1) *Not Applicable:* The EDR system does not deploy a sensor on the given platform and thus has no visibility.
(2) *None:* The EDR system deploys sensors on the given platform, but no data is available to show the event happened.
(3) *Telemetry:* The EDR system knows the event happened but is unsure if they are malicious.

(4) *General Behavior:* The EDR system knows the event happened and believes they are malicious. However, the system is unsure why and how the action was performed.
(5) *Tactic:* The EDR system knows the event happened and believes they are malicious. The system knows why the action was performed but is unsure how the action was performed.
(6) *Technique:* The EDR system knows the event happened and believes they are malicious. Additionally, the system knows why and how the action was performed.

In addition to the detection categories, modifiers provide more context about the detection. The modifiers include *delayed* and *config change*.

(1) *Delayed* means the alert appears significantly late compared to the time when the attack step happens.
(2) *Config change* means the alert shows up due to ad-hoc configuration modifications.

During the latest two evaluations, a new scenario was introduced to test the protection ability of EDR systems. The results of each step in the protection tests are categorized into one of the following protection categories.

(1) *Not Applicable:* The EDR system does not deploy a sensor on the given platform and thus has no visibility.
(2) *None:* The EDR system deploys sensors on the given platform but does not block the malicious behavior.
(3) *Blocked:* The EDR system successfully blocked the malicious behavior.

To quantify the detection performance of EDR systems, MITRE defines four metrics to summarize each EDR system's capabilities at a high level: *Visibility*, *Telemetry Coverage*, *Analytic Coverage* and *Detection Count*.

(1) *Telemetry Coverage* is the number of detected steps with the telemetry level detection. This is the minimum requirement for a step to be visible, as telemetry detection only confirms an event has happened but wouldn't trigger an alarm.
(2) *Analytic Coverage* is the number of detected steps with some contextual information like the intention and the approach taken. Since only detection above the telemetry level is reported as malicious behavior, the analytic coverage reflects a system's ability to detect threats from the available data.
(3) *Visibility* is the number of steps with at least a telemetry detection. Note that this metric counts the number of steps in the union of *Telemetry Coverage* and *Analytic Coverage*.
(4) *Detection Count* is the total number of detection made in the attack campaign. This number could be larger than the total steps, as multiple detections in different categories might be reported at a certain step.

2.2 Limitations of ATT&CK Evaluation

The MITRE evaluation has made significant contributions to establishing an evaluation standard for EDR solutions. However, many limitations still need to be addressed and improved upon.

2.2.1 Missing whole-graph analysis. The security field has been shifting from single-point detection to graph-based detection. The single-point detection can only detect a single step in an attack without providing an overview of the entire attack pattern. In contrast, whole graph-based detection utilizes contextual information to construct a comprehensive graph that depicts behavior and searches for threats. For modern endpoint APT defense, relying solely on single-point detection is inadequate for two reasons: Firstly, single-point detection is vulnerable to complex and sophisticated attacks that can evade traditional detection methods. Attackers can use multiple techniques to bypass single-point detection. Secondly, single-point detection focuses only on one aspect without considering contextual information, such as control and data flow, which limits perspective and could lead to false positive alarms.

Provenance graph-based detection [8, 26] overcomes these shortcomings by taking additional contextual information into account and obtaining a comprehensive view of the endpoint. Even if a single step is not identified as malicious from a single-point perspective, it can still be determined as part of the kill chain through control flow and data flow connections with other malicious behaviors. Moreover, using such correlations, malicious behaviors can be better distinguished from benign activities, reducing the number of false alarms. Most state-of-the-art endpoint detection work incorporates provenance graphs and their derivatives as part of their framework.

As discussed in the §2.1, MITRE uses a sequence of techniques to describe attack scenarios. However, the execution of kill chains in attack emulations hardly follows a linear pattern. Although such sequential representation emphasizes the detection performance on single steps, it obscures the causal and spatial aspects of the attack scenario. Fig. 1 shows an example of the attack graphs we constructed from an attack emulated in Wizard Spider+Sandworm (2022) evaluation. Without the graph, it's hard to understand the importance of each step in event correlation. For instance, `rundll32.exe` loads the downloaded malicious DLL file `adb.dll` to perform the following attack steps at step ⑧. Missing this step in the scope makes it hard to correlate the following attack steps with the previous setup steps. However, missing other less important steps, like `winword.exe` loading a malicious DLL file `VBDUI.DLL` at step ② does not affect the connectivity nor the causal relationship. With the help of the graph, we can investigate 1) whether the EDR systems can detect all crucial steps and 2) whether they can correlate events along the attack chain to reconstruct the attack chain and protect the system.

2.2.2 Lacking comprehensive interpretation. Another crucial part missing in MITRE evaluations is comprehensive interpretations of the results. Multiple companies claim they have achieved perfect or near-perfect performance in these evaluations. However, such claims contradict the results presented on the MITRE website. CrowdStrike [13] claims to have received 100% detection coverage across all 20 steps of the Carbanak+Fin7 evaluation. However, such a claim does not align with the evaluation results. Specifically, the '20 steps' are comprised of 174 substeps, where CrowdStrike failed to catch 22 out of 174 substeps and failed to generate alarms for 88 out of 152 visible steps. Other vendors have exhibited similar results.

Lacking clear interpretation leave room for EDR vendors to twerk the results, ironically going against the original intention of MITRE evaluation. Thus, it is necessary to add a comprehensive

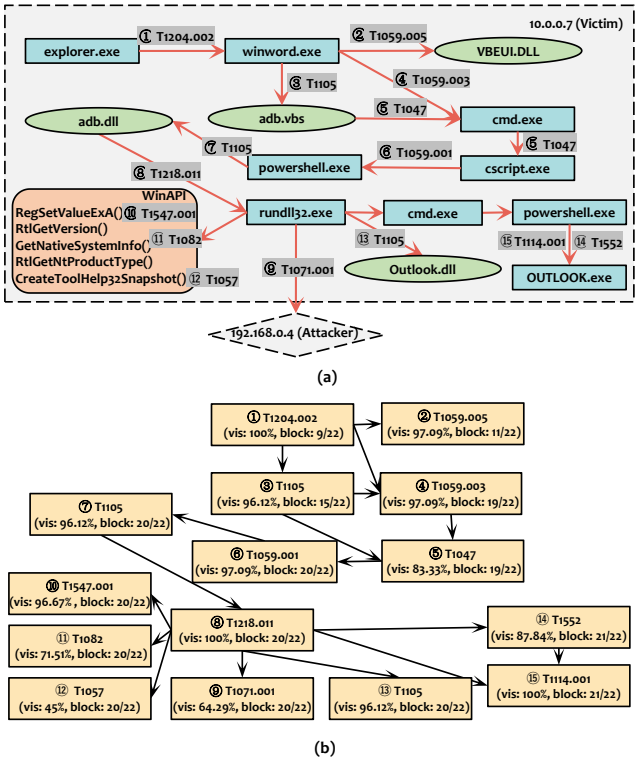


Figure 1: The attack graphs for scenario 1 in Wizard Spider+Sandworm (2022) evaluation. (a) The actual attack graph. The nodes are system entities like processes and files. The edges represent system events characterized by MITRE ATT&CK techniques IDs. The numbers denote the order of events. (b) The causal relationship attack graph. The nodes are attack steps characterized by MITRE ATT&CK techniques IDs. The edges represent causal relationships between attack steps. The nodes also contain the visibility of their corresponding techniques among all EDR systems and the number of EDR systems that blocked this attack before and at this step.

and objective interpretation on top of the MITRE Evaluation raw results.

2.2.3 Inconsistent evaluation framework. The MITRE Engenuity started the evaluation project in 2018. Since then, the evaluation approaches and terminology have been changing yearly, making it hard to compare the detection performance from different evaluations. For example, in the first APT3 evaluation, there are six detection categories, including *None*, *Telemetry*, *Indicator of Compromise*, *Enrichment*, *General Behavior*, and *Specific Behavior*. In the most recent Wizard Spider+Sandworm (2022) evaluation, there are five detection categories, including *None*, *Telemetry*, *General*, *Tactic*, and *Technique*. Although *None* and *Telemetry* remain the same, MITRE didn't map the rest of the detection categories. Besides, MITRE has used several versions of detection modifiers and even changed the definition of performance metrics over the years. In the most recent Wizard Spider+Sandworm (2022) evaluation, MITRE changed how detection numbers are counted. Instead of

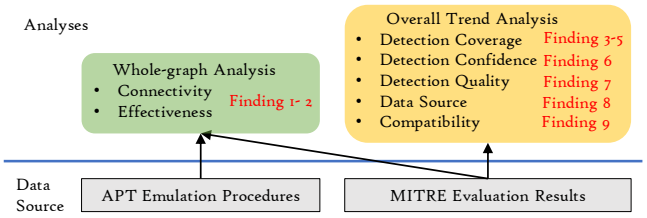


Figure 2: An overview of our analysis methodologies.

counting multiple detections on a single step, MITRE only recorded the detection with the most contextual information. In this way, visibility is the sum of telemetry and analytic coverage. The detection count metric is deleted since it is always the same as visibility.

Thus, in this paper, to bridge the gap between direct results and insight, we aim to provide a comprehensive interpretation of MITRE Engenuity Evaluation results. We also try to extract the consistent aspects from the different terminology used in each year's evaluation to establish a compatible interpretation framework to compare evaluation results from different years.

3 INTERPRETATION METHODOLOGY

3.1 Overview

We start this section by introducing the dataset in §3.2. Then, we elucidate two approaches to analyze this dataset: (1) whole-graph analysis and (2) overall statistical and trend analysis. Fig. 2 present an overview of our analyses. In the whole-graph analysis, we studied techniques provided in the evaluation results and APT emulation procedures published by MITRE Center for Threat-Informed Defense [18] to construct several causal relationship attack graphs. Via connectivity analysis and effectiveness analysis of the causal relationship attack graphs, we investigate the EDR systems' attack graph-level correlation and reconstruction capabilities. In the overall trend analysis, we investigate the detection performance of EDR systems on various techniques through several perspectives over the years. We aim to provide insights into the strengths and areas requiring enhancement within intrusion detection. We present the detailed methodologies in §3.3 and §3.4, respectively.

3.2 Dataset

The MITRE Engenuity has conducted four evaluations so far. Each evaluation selects one or two Advanced Persistent Threat (APT) groups and emulates several attack scenarios using the APT groups' toolkits. In total, attack scenarios consist of hundreds of steps involving dozens of techniques. For every attack scenario, MITRE Engenuity will conduct a detection test and a protection test. In detection tests, the attack kill chain will be executed without intervention. MITRE will assess whether every attack step is detected and the extent of contextual information provided during detection. While in protection tests, the defense team can intervene and stop the consequential steps in the kill chain. In this way, MITRE can assess whether an attack is contained or blocked and at which step. Therefore, the published dataset presents the results in a step-wise manner. For the detection results, data entries specify the following information about an attack step: (1) the MITRE ATT&CK technique that corresponds to the step, (2) whether the step is detected

Table 1: MITRE Engenuity Dataset Summary

Round of Evaluation	Participants	Steps	Techniques	# of Detection Made
APT3 (2018)	12	136	51	1970
APT29 (2019)	21	134	53	3982
Carbanak+FIN7 (2020)	29	174	46	7350
Wizard Spider+Sandworm (2022)	30	109	46	3098
Total	37	N/A	82	16.4k

and how much contextual information is provided by the EDR system, (3) the data sources associated with detection, and (4) other miscellaneous information. The protection results also contain the MITRE ATT&CK technique corresponding to the step. Besides, instead of the detection-related information, the results show what kind of protection is triggered at each step. The definition of detection and protection categories is detailed in §2.1. Table 1 outlines detailed information about the dataset associated with each campaign. Overall, we analyzed 16.4k detection results from all vendors in all published evaluation results.

3.3 Whole-graph Analysis

3.3.1 Attack Graph Construction. Due to the importance of provenance graph-based detection capabilities, we create a causal relationship graph model for each scenario to replace the sequential layout of MITRE evaluation results. A causal relationship attack graph is a directed graph in which the node represents an attack step in the kill chain, and the edge denotes the causal relationship between two attack steps. Constructing a causal relationship attack graph model involves nodes construction and edges construction.

Since MITRE only describes each step in terms of techniques, our initial step is to thoroughly examine each step's corresponding procedure, which is subsequently classified into descriptive and causal categories. The descriptive techniques solely describe specific features of a step without establishing any causal connections with other entities (e.g., *Encrypted Channel*). In contrast, the causal techniques interact with other entities, such as creating a process or writing to a file, thereby establishing causal relationships with other steps (e.g., *Ingress Tool Transfer*). All steps corresponding to causal techniques become the nodes in our causal relationship attack graph.

After classifying each step, we examine the subject and object of each step to establish causal relationships. When we connect the subject to the object, the edges are constructed in the graph. We generally consider two kinds of causal relationships: control flow and data flow. Firstly, the control flow creates causal relationships via process creation. If an attack step is performed by a process created in a previous step, then the two steps establish a control flow causal relationship. Secondly, the data flow creates causal relationships via communication over files. If an attack step reads a file written in a previous step, the two steps establish a data flow causal relationship. Fig. 1(b) and 5(b) in the Appendix show two examples of causal relationship attack graphs.

3.3.2 Attack Graph Analysis. After constructing a causal relationship attack graph, we analyze EDR systems' detection and protection performance from the attack graph perspective. We aim to answer the following two questions: (1) whether the EDR systems can fully reconstruct the attack kill chain; (2) whether the EDR systems can effectively aggregate behaviors along the kill chain to understand its severity. We answer the first question by conducting a connectivity analysis on the causal relationship attack graphs constructed from the detection results. As for the second question, we analyze the protection performance to examine the EDR systems' effectiveness. An effective EDR system should detect and respond to threats at the appropriate time. We study several attack cases and investigate when EDR systems block the kill chain.

Connectivity Analysis: We examine the kill chain visibility by counting the connected components in the causal relationship attack graph and comparing them with the ground truth. Suppose a campaign involves attacks on three individual hosts, leading to three separate kill chains among those hosts. If the number of connected components on the graph is more than three, one or more kill chains are broken into multiple small segments. If the number is less than three, at least one kill chain is completely missing in the detection. If the number equals three, we check if the three segments match the ground truth. In this way, we use the connected components as a metric to evaluate EDR systems' attack reconstruction capability.

Effectiveness Analysis: We examine which step a blockage is triggered given each vendor's detection results to analyze how effectively the EDR systems use the graph information. We assume EDR systems are knowledgeable about the maliciousness of different behaviors, and they would block the attack once the severity of existing behaviors accumulates to a certain threshold. We select a few attacks to perform case studies. For each case, we manually determine a step on the attack kill chain when the malicious intention is evident as the baseline. Then, for each EDR system, we compare at which step the attack is blocked with the baseline. If the attack is blocked earlier than the baseline, it suggests the EDR system adopts an aggressive strategy in defense response. In this case, benign behaviors could be incorrectly classified as malicious, leading to unpredictable problems. If the attack is blocked later than the baseline, it suggests the EDR system cannot react to threats in time. Such delay could allow the attacks to happen unhindered.

3.4 Overall Trend Analysis

Besides analyzing the evaluation results on the attack graph, we also investigate the detection and protection performance of individual techniques and EDR systems over the years.

3.4.1 Detection Coverage. MITRE presents the evaluation from the vendors' perspective. Each vendor's performance is reflected by its analytic coverage, telemetry coverage, and visibility on all attack steps. We want to investigate EDR systems' performance from another perspective: how well all vendors can detect each technique.

We analyze the detection coverage of a technique from two perspectives: visibility and analytic coverage. Like MITRE's metrics, we determine the visibility of a technique by calculating the ratio of EDR systems aware of the corresponding behaviors. We determine the analytic coverage of a technique by calculating the percentage of the vendors that successfully detected the corresponding step.

(1) *Visibility*

$$V(x) = \frac{S_v(x)}{S_t(x)}$$

Where x is a specific EDR system or a specific technique. $V(x)$ is the visibility score of the given x. $S_v(x)$ is the number of visible substeps of the given x, and $S_t(x)$ is the total substeps related to the given x.

(2) *Analytic Coverage*

$$A(x) = \frac{S_a(x)}{S_v(x)}$$

Where $A(x)$ is the analytics score of the given x, $S_a(x)$ is the number of detections beyond the telemetry level of the given x, and $S_v(x)$ is the number of visible substeps of the given x.

3.4.2 Detection Confidence. Although the three metrics adopted from MITRE evaluation provide helpful information, some aspects are missing. Specifically, those metrics do not take different detection categories into account. In other words, a system reporting all malicious behaviors in the general behavior level would receive identical scores as a system reporting all malicious behaviors in the technique level. We propose an additional metric: *confidence* to address this issue.

Confidence is a weighted score calculated by multiplying the percentage of detection made from different detection methods by the corresponding weight multiplier. A high confidence score suggests more details about the malicious behavior are provided.

$$C(x) = \frac{4D_{te}(x) + 3D_{ta}(x) + 2D_{ge}(x) + D_{tel}(x)}{4D_v(x)}$$

Where $C(x)$ is the confidence score of a given technique or EDR system x. $D_{te}(x)$ is the number of technique detection of the given x, D_{ta} the number of tactic detection of the given x, D_{ge} the number of general behavior detection of the given x, D_{tel} the number of telemetry detection of the given x, and D_v the total visible substeps of the given x. The multiplier associated with each variable indicates the granularity of the detection results, with four being the most detailed and one being the most general. Since MITRE evaluation provides four levels of granularity for detection results, we intuitively use 1 through 4 as the multipliers. They can be further adjusted if more detailed data is available.

3.4.3 Detection Quality. Another aspect missing from MITRE-provided metrics is the negative modifiers. A system reporting all alarms with significant delay or configuration change receives identical scores as a system reporting all alarms with no delay and no configuration change. To examine the presence of modifiers quantitatively, we propose a *quality* metric for techniques and EDR systems. For a technique, the quality score is the ratio of visible substeps without negative modifiers to the total visible substeps. A high-quality score implies low detection latency and adequate out-of-box usability.

$$Q(x) = \frac{S_m(x)}{S_v(x)}$$

Where $S_m(x)$ is the number of visible substeps without negative modifiers of a given technique or EDR system x, and $S_v(x)$ is the total visible substeps of the given x. We treat all the negative modifiers equally since they are all related to manual adjustments or analyses.

3.4.4 Data Source. Besides the quantitative analysis specified above, we investigate the data sources used in evaluations via a rather qualitative approach. Specifically, we compare the data sources used in each year's evaluation to examine the scope of data sources used in EDR systems. We also discuss the frequency of a data source used in each evaluation to investigate the importance of the data source.

3.4.5 Compatibility. We investigate EDR systems' compatibility from two perspectives: availability and performance. We examine the availability by calculating the ratio of EDR systems that support a given platform. Since MITRE Engenuity evaluations only involved Windows and Linux platforms so far, we will focus on the availability of these two platforms. Besides, we compare the detection performance on different platforms.

4 WHOLE-GRAPH ANALYSIS

Since whole-graph analysis is specific to the attack scenarios, we use attack scenarios in the Wizard Spider+Sandworm (2022) evaluation as the cases to perform our whole-graph analysis. We constructed casual relationship attack graphs at the procedure level for all attack scenarios in Wizard Spider+Sandworm Evaluation. Fig. 1 and 5 in the Appendix present two examples of constructing the causal relationship attack graphs.

4.1 Connectivity Analysis

We analyze the attack graph connectivity by calculating the number of connected components in the casual relationship attack graph generated from the visible steps of each vendor and comparing it with the ground truth. There are six hosts involved throughout the attack emulation. Thus, there should be six segments. One of the six hosts runs under the Linux environment, and the other five are under the Windows environment. 22 vendors support Linux environment data collection and detection out of 30 participants. Therefore, we divide the vendors into two groups according to their Linux platform compatibility. Of the 22 vendors that support the Linux platform, three have more than six segments (Rapid7 has 13, Cisco has 10, and Cylance has 11). Of the eight vendors that don't support the Linux platform, two have more than five segments (Deep Instinct has nine, and ReaQta has six). 25 out of 30 (83.3%) participants can obtain a visibly connected subgraph containing all attack steps. Thus, we conclude that most vendors can see the connection between attack steps on a graph level.

4.2 Effectiveness Analysis

We analyze all protection evaluation scenarios and discuss two as case studies to see how effectively the EDR systems use the graph

Table 2: Summary of Protection Test Results

Test	# of Blockage	# of Participants	Protection Rate
1	21	22	95.5%
2	21	22	95.5%
3	16	22	72.7%
4	12	22	54.5%
5	15	22	68.2%
6	20	22	90.9%
7	9	17	52.9%
8	20	22	90.9%
9	18	22	81.8%

information. Table 2 summarizes the results of nine protection tests. 22 vendors participated in the protection tests. Test 7 is conducted on Linux. Since five participants didn't support the Linux platform, only 17 were in Test 7. Fig. 1 and 5 in the appendix presents the two cases we will discuss in detail.

Scenario 1: Emotet Initial Compromise, Persistence, and Collection.

Fig. 1 shows a detailed attack graph of this scenario. In this scenario, the adversary sent a Word document over email, which contained obfuscated VBA macros that downloaded and executed a malicious DLL based on the malware Emotet. The malicious DLL then established a command and control (C&C) session with the adversary server. Besides, it achieved persistence by modifying the registry via the WinAPI function `RegSetValueExA()`. Later, the malicious DLL collected process information by calling the WinAPI functions `CreateToolhelp32Snapshot()` and `Process32First()`. Finally, it downloaded another malicious DLL to search for credentials in Outlook.

21 out of 22 EDR systems blocked this attack at different steps. Most blockages happen at the first step when the Explorer executes the Word document. This behavior is already pretty suspicious, as it downloaded an untrusted file. In the following steps, Word downloaded a malicious DLL file and a malicious VBS file and then executed it. The malicious intention is evident at this step, as this is a typical download and execution behavior. However, not all EDR systems responded to it: 19 EDR systems blocked the process, while three EDR systems either waited until later to block or didn't react. We checked their connectivity and found they could see the connected attack steps corresponding to this protection test. Although they have reasonable detection performance on single steps, these EDR systems failed to chain steps together to better understand the kill chain to provide appropriate protection.

Scenario 2: TrickBot Execution, Discovery, and Kerberoasting.

Fig. 5 in the Appendix shows a detailed attack graph of this scenario. In this scenario, the adversary authenticated into the victim's host using stolen credentials from scenario 1. Then, the adversary downloaded and executed a malicious EXE derived from TrickBot. The malicious EXE first established a C&C session with an adversary-controlled server. Then, it collected various system information by executing shell commands. Finally, it downloaded a tool called `rubeus` to perform Kerberoasting [14], which could steal encrypted credentials.

21 out of 22 EDR systems blocked this attack at different steps. In this scenario, the adversary connected to the target via RDP protocol as the first step. Looking at this step alone, it could be normal behavior. However, in the latter steps, the adversary downloaded a malicious file and executed it to establish a communication channel with the C&C center. The malicious intention is evident at this step as it downloaded and executed an unknown file and established a suspicious outward connection. Only half of the EDR systems decided to block the process at steps 3 and 4. The rest of the EDR systems block the process when collecting system information in the later steps.

Although scenarios 1 and 2 had the same protection rate eventually, there is a noticeable delay in scenario 2 compared to scenario 1. One reason could be the difference in step visibility. As shown in Fig. 1(b) and 5(b), early steps in scenario 1 had better visibility than early steps in scenario 2. The third step in scenario 2 only had 64.29% visibility, making it hard for EDR systems to gather enough information and react.

Finding 1: Attack graph level correlation capabilities are necessary to achieve good defense because isolated single steps cannot provide enough confidence for EDR systems to respond.

The steps in Tests 3, 4, 5, and 6 happened as a connected kill chain on the same host in the detection test but are isolated into different test scenarios in the protection tests. This gives us a chance to investigate how isolated scenarios can affect defense. Since isolated scenarios contain fewer steps for correlating, the defense and response decisions primarily rely on single-step detection. We observed a significant drop in protection rate in tests 3, 4, and 5 as shown in Table 2. Test 4 only contained two steps that dumped system information (C disk and the registry) and received the lowest protection rate. Although dumping the entire C disk and the registry seems suspicious, such behaviors alone are usually not malicious enough to be escalated to alarms. Admittedly, we observed many cases in which EDR systems take action as soon as a suspicious file is downloaded and executed. Still, such a download and execution pattern wouldn't work well against file-less attacks, living-off-the-land attacks, and other evasion techniques.

Furthermore, some evidence shows EDR systems with poor performance didn't have graph-level correlation capabilities. For example, the detection screenshots from vendors like Deep Instinct didn't present any graph-level information along with the detection. In contrast, the detection screenshots from vendors like Sentinel One complement the detection with kill chain information on a graph.

Finding 2: Although some EDR systems demonstrated good attack graph level correlation capabilities, we still identified three practical problems: delay in protection, lack of protection, and lack of cross-host correlation capability.

Delay in protection problem exists ubiquitously in all scenarios. In the cases we analyzed, the adversary will log into the target and download a payload. More than half of the protection happens here, but the rest would happen either after the adversary had done some malicious behaviors or not at all. We checked their visibility. Most of them can see a connected attack chain. Those EDR systems require a longer kill chain to accumulate confidence before blocking a process. Such a mechanism prevents them from reacting quickly to threats. Sometimes, this even prevents them from reacting at all.

Lack of protection problems would still occur when some stealthy steps are applied. Besides requiring a longer chain to reach their confidence level, some EDR systems are susceptible to attack evasion techniques. For example, Test 3 modified the registry to achieve persistence. Tests 4 and 5 mimic system administrators to dump system information and modify system configurations. These tests applied more stealthy and sophisticated approaches than other tests, thus receiving a relatively low protection rate.

Attack graph level correlation should be applied on individual hosts and across hosts. In this evaluation, the adversary used the same tools and adopted similar attack patterns on different hosts in tests 2 and 3, respectively. Given test 2 happened before test 3, the protection performance in test 3 is not better than test 2. It suggests the EDR systems could not learn from the happened attacks to react to similar attacks in the future. Furthermore, no evidence shows that detection and response mechanisms use information across hosts to improve defensive performance.

5 OVERALL TREND ANALYSIS

In this section, we examine several perspectives in all available datasets from MITRE Engenuity to investigate the paradigm of attacks and defenses in a real-world setting. We mainly analyze the results from more recent evaluations, especially the Carbanak+Fin7 (2020) and the Wizard Spider+Sandworm (2022) evaluations because they included APT emulations performed on multiple operating systems and they used a more established taxonomy to describe evaluation results compared to the previous evaluations.

5.1 Detection Coverage

We calculate the visibility and analytic coverage scores for individual techniques across the EDR systems participating in the evaluations. The distribution of visibility and analytic coverage scores from the technique perspective are shown in Fig. 3. We also calculate the visibility and analytic coverage scores for individual EDR systems as shown in Fig. 4. Then, we use these two metrics to analyze the overall trend of detection coverage. Moreover, we select a few EDR systems and techniques that receive excellent or very low scores and try to analyze the reasons behind them.

5.1.1 Visibility. As shown in Fig. 3 and 4, the distributions of technique visibility scores and vendor visibility scores are mostly skewed to the left. The median of technique visibility distribution from Wizard Spider+Sandworm (2022) evaluation is around 95%, which means half of the attack steps can be seen by at least 95% of the EDR systems in evaluations. The median vendor visibility distribution from the same evaluation is around 85%, suggesting half of the EDR systems can see more than 85% of the attack steps. Comparing the visibility score distribution in the three evaluations, we see an obvious improvement in the median and the lowest score for both techniques and vendors over the years.

Two techniques in Command and Control (C&C) receive low scores across the EDR systems in the Carbanak+Fin7 evaluation. Specifically, only around 40% of the EDR systems can record *Encrypted Channel* or *Application Layer Protocol* communications. In this evaluation, such communications include transmitting data over SSH protocol, MSSQL transactions, and HTTPS protocol. Meanwhile, the visibility of other C&C techniques, which establish the connection via TCP, are all above 80% among the EDR systems. Based on this observation, we conclude most EDR systems selectively collect network traffic data on the transport layer and ignore the application layer protocols.

Surprisingly, 15 techniques achieved perfect coverage among all EDR systems in the Wizard Spider+Sandworm evaluation, including six discovery techniques, three Defensive Evasion techniques, and a few techniques in other tactics. Those techniques involve manipulating or investigating system configurations and services, which implies all EDR systems have no trouble monitoring system configurations and services. In addition, four out of six techniques in the Execution Tactic receive a visibility score above 90% in the Carbanak+Fin7 evaluation. Those techniques involve a process loading certain libraries or executing specific commands. We conclude that all EDR systems emphasize monitoring process loading and execution so that the techniques related to execution achieve the best visibility among all techniques.

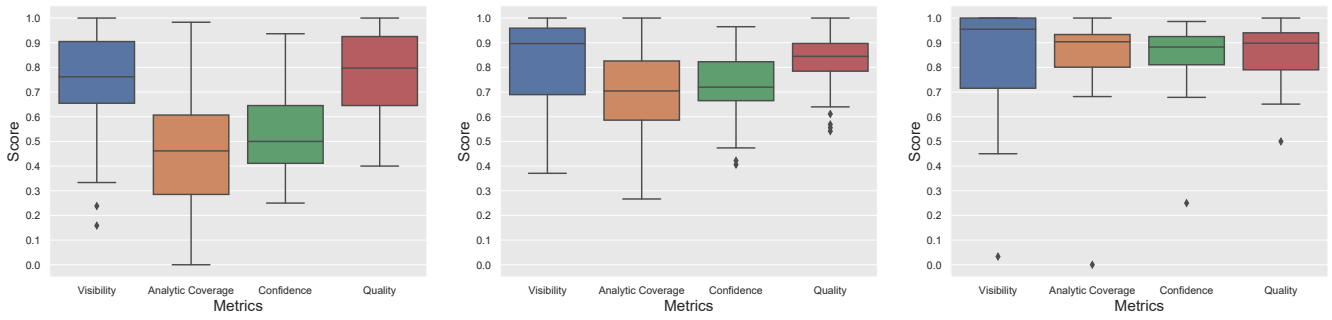
SentinelOne achieves 100% visibility on all techniques used in the Carbanak+Fin7 evaluation, while only around 50% of the techniques are visible to AhnLab in the same evaluation. AhnLab hardly monitors the file system, as most of the techniques associated with *Collection* and *Credential Access* remain invisible to AhnLab. Furthermore, AhnLab does not monitor network activities except for a few file download behaviors and some TCP connections. It is particularly interesting that AhnLab can see and even raise alarms on some file download behaviors by Powershell but remain completely blind to the rest of the file download behaviors by Powershell from the same IP.

We observe changes and continuities by comparing the visibility score between the Carbanak+Fin7 and Wizard Spider+Sandworm evaluations. Most EDR systems obtain higher visibility scores in the latter evaluation, especially AhnLab, whose visibility score has a huge boost from 0.517 to 0.761. This implies the entire industry has improved in making various behaviors visible. The visibility difference in techniques is interesting. The two evaluations have about the same average visibility scores (about 80%). However, some techniques like *Archive Collected Data* have a perfect visibility score in the Carbanak+Fin7 evaluation but only receives a 0.033 visibility score in the Wizard Spider+Sandworm evaluation, which means it's only visible to one out of 30 EDR systems. The dramatic discrepancy suggests technique is not an appropriate unit for detection coverage since the same technique could be implemented with totally different procedures and consequently require distinct detection capabilities.

Finding 3: Most EDR systems have good data collection capability, and this capability is improving every year. In the most recent Wizard Spider+Sandworm (2022) evaluation, 75% of the EDR systems can identify more than 80% of the attack steps.

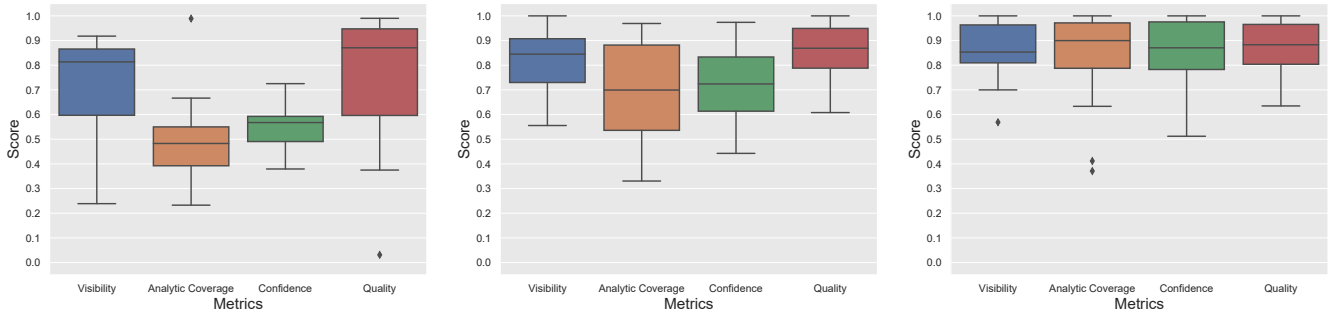
Finding 4: Large discrepancies in visibility for the same technique in different evaluations suggest that techniques are still too coarse-grained for detection coverage. A more fine-grained unit, such as generalized technique implementations, is needed.

5.1.2 Analytic Coverage. As shown in Fig. 3 and 4, analytic coverage has significantly improved over the years. In the most recent Wizard Spider+Sandworm (2022) evaluation, 50% of visible attack



(a) APT29 (2019) (b) Carbanak+Fin7 (2020) (c) Wizard Spider+Sandworm (2022)

Figure 3: Technique perspective score distribution of each metric in different evaluations. The metrics are visibility (blue), analytic coverage (orange), confidence (green), and quality (red) from left to right, respectively.



(a) APT29 (2019) (b) Carbanak+Fin7 (2020) (c) Wizard Spider+Sandworm (2022)

Figure 4: Vendor perspective score distribution of each metric in different evaluations. The metrics are visibility (blue), analytic coverage (orange), confidence (green), and quality (red) from left to right, respectively.

steps would trigger alarms on more than 90% of the EDR systems, and 50% of the EDR systems would generate alarms for at least 90% of the visible attack steps. Unlike the distribution of visibility, analytic coverage exhibits a nearly symmetric distribution.

The techniques receiving low analytic scores are mainly in C&C, Exfiltration, and Collection tactics. Not surprisingly, *Archive Collected Data* has an analytic coverage score of 0 given its low visibility score, which means only telemetry detection is made for this technique in the Wizard Spider+Sandworm evaluation. Other techniques like *Exfiltration Over C&C Channel* and several C&C techniques also received below 50% analytic coverage scores. Although such behaviors might be indistinguishable from everyday benign behaviors, they could still be identified as a part of a complete attack chain. Thus, detecting these behaviors challenges EDR systems' capabilities of assembling attack chains from scattered individual events. Low analytic coverage scores on those techniques suggest all EDR systems have room for improvement in their event correlation capability.

Moreover, the techniques receiving high analytic coverage scores mostly fall into categories of Defense Evasion and Credential Access, including *OS Credential Dumping*, *Inhibit System Recovery* and *Process Injection*. Such techniques are easy to separate from normal behaviors as they carry malicious intentions conspicuously.

Comparing the analytic coverage scores between the Carbanak+Fin7 and Wizard Spider+Sandworm evaluations, the analytic coverage scores have a significant improvement for all techniques as the average increases from 69.8% to 85.5% and the standard deviation decreases from 16% to 8%. Techniques in C&C, Exfiltration, and Collection tactics that receive less than 60% analytic coverage scores in the Carbanak+Fin7 evaluation mostly have higher than 70% analytic coverage scores in the Wizard Spider+Sandworm evaluation.

Finding 5: Although EDR systems' ability to determine malicious behaviors improves over time, they struggle to detect 'living-off-the-land' threats. Throughout all evaluations, several techniques, including *Encrypted Channel*, *Exfiltration Over C&C Channel*, and *Archive Collected Data*, etc., consistently had worse detection rates. It is hard for EDR systems without event correlation capabilities to discern whether such individual steps are malicious at the system provenance level.

Moreover, we divide EDR systems and techniques into four quadrants according to their visibility and analytic coverage scores for comparison. As shown in Fig. 6 and 7 in the Appendix.

For EDR systems, the top performers like Sentinel One and Palo Alto Networks at the top right corner are excellent at visibility and detection, and the bottom performers like AhnLab at the bottom left corner need to catch up with both capabilities. Interestingly, there are some ramifications among the EDR systems in the middle. The five EDR systems in the top left quadrant, like FireEye, tend to focus on detection capability. FireEye has way above average analytics score (90.3% vs. 69.8%) but slightly below average visibility score (78.2% vs. 80.8%). On the contrary, the six EDR systems in the bottom right quadrant, like CrowdStrike, tend to put more

emphasis on visibility capability. CrowdStrike has a higher-than-average visibility score (87.4% vs. 80.8%) but a lower-than-average analytic coverage score (46.2% vs. 69.8%).

For techniques, 42 out of 63 falls into the top right quadrant and the bottom left quadrant, meaning they have comparable visibility and analytic coverage scores. For instance, *Network Share Discovery* has an excellent visibility score (100%) and analytic coverage score (90.1%), while *Exfiltration Over C&C Channel* has a poor visibility score (43.7%) and analytic coverage score (22.6%). Six out of 63 techniques fall into the top left quadrant. Those techniques receive below-average visibility scores but above-average analytic coverage scores. It implies it is hard for EDR systems to collect related data, but they can be easily identified as malicious behaviors once visible. For example, *Access Token Manipulation* receives only a 42.9% visibility score but an 80.6% analytic coverage score. Such techniques often involve some defense evasion intentions, or it might be expensive to collect related data, making them naturally stealthy. 15 out of 63 techniques fall into the bottom right quadrant. Those techniques receive above-average visibility scores but below-average analytic coverage scores. It suggests they are easily visible to EDR systems, but it is hard to associate them with malicious behaviors. For example, *Email Collection* has a 95.2% visibility score but only a 37.5% analytic coverage score. Such behaviors are not stealthy but blend in with benign behaviors a normal user would perform.

To sum up, 18 out of 29 EDR systems have comparable visibility and analytic coverage scores. In comparison, five of the rest focus more on their detection capability, and six EDR systems lean towards improving their visibility. 42 out of 63 techniques have comparable visibility scores and analytic coverage scores. Among the remaining 21 techniques, six techniques are hardly visible but easy to detect once they are visible. The other 15 techniques are easily visible but hard to identify as malicious behaviors.

5.2 Detection Confidence

The layout of confidence scores looks similar to analytic coverage scores since they reflect the soundness of detection results. However, we propose the confidence metric in addition to the analytic coverage metric because the confidence metric takes different detection levels above telemetry into account. As a weighted average, the confidence score indicates the overall detection level made by an EDR system or against a technique. Along the spectrum of confidence scores, a 25% confidence score means the detection level is at telemetry on average; a 50% confidence score means the detection is at General Behavior on average; a 75% and 100% confidence score means the detection level is at Tactic and Technique, respectively. For instance, *Input Capture* has a perfect analytic coverage score, while it only has a 61.7% confidence score in Carbanak+Fin7 (2020) evaluation, suggesting most of the detection levels are between General Behavior and Tactic. Other Credential Access techniques like *Unsecured Credentials and Credentials from Password Stores* also have discrepancies in analytic coverage and confidence.

Interestingly, the techniques that receive the highest and lowest confidence scores are all in Credential Access or Discovery tactics. *OS Credential Dumping* and *Network Share Discovery* receive above 90% confidence scores, whereas *Credentials from Password Stores* and *Permission Group Discovery* obtain below 30% confidence scores. All EDR systems can identify malicious behaviors related to *OS Credential Dumping* and *Network Share Discovery* and complement alarms with additional context like motivations and techniques used. On the contrary, EDR systems struggle to set off alarms for malicious behaviors related to *Credentials from Password Stores* and *Permission Group Discovery*, let alone providing additional context.

The technique *Web Service* has a significantly lower confidence score than its analytic coverage score. This suggests the alarms on *Web Service* fail to provide detailed contexts, like the motivation of such behavior or the specific technique used. Such vague alarms usually require system administrators to spend more time investigating and responding. On the other hand, techniques like *Exfiltration Over C&C Channel* remarkably higher confidence scores than its analytic coverage score, which implies detection on *Exfiltration Over C&C Channel* comes with a satisfying amount of details although the number of generated alarms is relatively low.

For EDR systems, the confidence scores and analytic coverage scores also share the general trend but with a few outliers. We calculate the confidence-analytic difference across all EDR systems participating in Carbanak+Fin7 (2020) evaluation to further investigate their detection confidence. CyCraft has a notably lower confidence score than the analytic coverage score, suggesting CyCraft puts more emphasis on detection coverage than other contexts. On the contrary, EDR systems like Sophos and AhnLab have much higher confidence scores than their analytic coverage score, which suggests they value the context provided in detection more than the coverage. Top performers like Palo Alto Networks, Sentinel One, and CheckPoint have about the same confidence score and analytic coverage score, which implies their detection has satisfying coverage and abundant details.

Comparing the distribution of confidence scores over the three years in Fig. 3 and 4, the technique median confidence score improved from 50% in APT29 to 72% in Carbanak+Fin7, and finally to 88.25% in Wizard Spider+Sandworm. The confidence score distributions exhibited a decrease in range while shifting towards the higher end. This implies a significant enhancement in the level of detail with the detection. In the APT29 (2019) evaluation, only 50% of the attack steps can be detected at the General Behavior level or above by all EDR systems on average; however, in the Wizard Spider+Sandworm (2022) evaluation, 75% of the attack steps can be detected at the Technique level on average.

Finding 6: Different amounts of details in alarms reflect EDR systems' detection confidence. Throughout the five evaluations, EDR systems are less confident to trigger alarms on techniques that widely exist in everyday activities such as *Email Collection*, *Exfiltration Over C&C Channel*, and *Ingress Tool Transfer*. Thus, EDR systems tend to provide less contextual information, like their roles in the attack chain. On the contrary, EDR systems are more confident in detecting typical malicious behaviors like *OS Credential Dumping* and *Network Sniffing*. EDR systems' overall detection confidence has improved remarkably, as shown in Fig. 3 and 4.

Table 3: Data Sources in MITRE Evaluations

Campaign	# of Data Source	Top 5 Data Sources
APT29 (2019)	9	File, Command, Process, Script, Network Traffic
Carbanak+FIN7 (2020)	25	Process, File, Network Traffic, Script, OS API Execution
Wizard Spider+Sandworm (2022)	41	Process, File, Network Traffic, OS API Execution, Logon Session

5.3 Detection Quality

We calculated the quality metric for techniques and EDR systems in the recent three evaluations, as shown in Fig. 3 and 4, respectively. The detection quality score has been increasing over the years. *Credentials from Password Stores* receive the lowest quality score (54.2%) in the Carbanak+Fin7 evaluation, which suggests significant delay and manual efforts are involved. In this scenario, the credentials stored in the Chrome web browser are accessed via a malicious tool. This technique has a fairly low Visibility score and analytic coverage score. Even when it is detected, it usually requires human analysis. Other techniques with low quality scores (below 60%) include *Inter-Process Communication*, *Credentials from Password Store* and *Data from Local System*, which require modifying detection policies depending on the local environments. We suspect failures in detecting this technique are also related to systems' weak ability to link individual events to an attack chain. While accessing the credentials stored in the browser itself doesn't seem very suspicious, downloading an unknown tool and using it to access credentials makes it very suspicious.

OpenText and DeepInstinct receive the lowest quality score (around 60%) among the EDR systems, while some EDR systems like SentinelOne, ReaQta, and CyCraft obtain close perfect quality scores in the Carbanak+Fin7 evaluation. The score differences imply EDR systems' various self-adapting abilities. Systems with high quality scores can work effectively in different environments without much human intervention, while systems with low quality scores require a lot of manual tuning and analysis.

Finding 7: EDR systems often require extra manual effort to detect techniques that are closely integrated with local environments, such as *Credentials from Password Stores* and *Inter-Process Communication*. Only four tested systems in the Wizard Spider+Sandworm (2022) evaluation do not require extra effort to detect such techniques.

5.4 Data Source

The number of distinct data sources has been changing over the years. No data source information is available in the APT3 evaluation (2018). Since the APT29 evaluation (2019), MITRE has started to collect data source information in the detection results. As shown in Table 3, nine distinct data sources are recorded in the APT29 evaluation results, whereas in the most recent Wizard Spider+Sandworm evaluation (2022), 41 different data sources are recorded. As the number of distinct data sources increases over the years, not only do the existing data sources become more specific, but some new data sources are also included in the data sources. At the same time, the taxonomy of data sources has been changing. In the APT29 and Wizard Spider+Sandworm evaluations, the data sources are recorded in *category: sub-category* format. In contrast, in the Carbanak+FIN7 evaluation, the data sources are recorded as *category* without further sub-categories. In the APT29 evaluation, the data sources are from the process, file, registry key, and network connection creations, as well as script and command line executions. In the Wizard Spider+Sandworm evaluation, network-related data sources include network connection creation, traffic content, and traffic flow. Besides, additional data sources, such as firewall metadata and network share access, are included. Our findings in §5.1 demonstrated such enrichment in data sources has a positive correlation with the improvement in the detection performance over the years.

Finding 8: Increasing complexity and variety of data sources suggest EDR systems can utilize extensive information from different dimensions. Although an increasing number of data sources are used, the top data sources remain unchanged. Process, file, network, scripts, and system calls/APIs are still the most fundamental and valuable data sources for EDR systems.

5.5 Compatibility

In the first two evaluations, APT3 and APT29, the target environments only contain Windows hosts; thus, all participants must support the Windows platform. Starting from the Carbanak+Fin7 evaluation in 2020, MITRE enrich the variety of target environments by including Linux servers as parts of the target system. In the Carbanak+Fin7 evaluation, 22 out of 29 participants supported the Linux platform. In the following Wizard Spider+Sandworm evaluation, 22 out of 30 participants supported the Linux platform.

In §4.2, we found EDR systems had a low protection rate against attacks on Linux. The attack on the Linux platform follows a similar pattern to the ones on Windows: uploading a payload and using it to establish C&C connections. Given the attacks have similar visibility and attack pattern, most vendors can protect against the attacks on Windows, but the attacks on Linux were only blocked by around half of the EDR systems.

Finding 9: Data collection and protection capability needs improvement on the Linux platform since around 25% of the evaluated EDR products don't support the Linux platform. For similar attack patterns, EDR products present worse protection results on Linux than on Windows.

6 RELATED WORK

6.1 Endpoint Detection and Response (EDR)

An increasing number of researches on endpoint security solutions have been conducted to improve APT defense methods and forensics technologies on various platforms. EDR frameworks like HOLMES [32], Poirot [31], MORSE[25], and others [4, 21, 23, 24, 28, 38–43] aim to improve defense on Windows and Linux operating system, while other methods like RiskRanker [20] and E-EMD [29] target security on mobile and cloud platforms, respectively. However, they all use statistical measurements like false positive and

true positive, precision, recall, accuracy, and F-Score, to describe the detection performance. They also include CPU and memory usage as measurements for overhead. However, it is hard to compare the measurements from different works due to the different datasets and hardware used to carry out the measurements. Surveys on the EDR systems [5, 27, 44] mainly focus on the methodology and dataset used but pay little attention to EDR evaluation.

In efforts to improve security evaluations, researchers have been studying the evaluation flaws in existing security works. For instance, Van Der Kouwe et al.[37] identifies a list of common benchmarking mistakes and indicates that benchmarking flaws exist widely in system security papers published in top venues, which suggests the necessity of standardizing security benchmarks. Following those papers' insights and suggestions, we propose our evaluation and interpretation framework for system security work in academia.

6.2 Security Benchmark

Most existing works mainly focus on generating representative data sets since quality security data sets are scarce. As early as 1998, DARPA launched its intrusion detection evaluation [15] in collaboration with Lincoln Lab at MIT. Zuech et al. [45] tried to generate network-based data sets for evaluating network intrusion detection systems (NIDS); Divekar et al. [17] modified existing network-based data sets to improve training performance in anomaly-based NIDS; Almakhdhub et al. [3] targets benchmarking Internet of Things (IoT) devices. Additionally, the data set from the DARPA Transparent Computing program [16] has been used widely in recent security work. Still, data from only two out of five attack campaigns are publicly available, and thus the attack scenarios are minimal. Although such data sets help mitigate the deficit in security evaluation data, they do not provide extensive methodologies for interpreting the results.

Some other work aims to improve the explainability of evaluations. Hao et al. [22] and Mendes et al. [30] designed methodologies to obtain more explainable evaluation results for static application security testing (SAST) tools and web serving systems, respectively. However, their methods are specific to the targeting tools or systems. Thus, it is hard to expand the methodologies to other security fields.

Recently, some security studies have used the knowledge base built by MITRE ATT&CK. Choi et al. [10] used the tactic, technique, and procedure (TTP) proposed by MITRE ATT&CK to generate attack sequences. On the other hand, Outkin et al. [35] use attacks emulated in MITRE ATT&CK evaluation as the attack models to discuss defender policy and resource allocation. Although they used MITRE ATT&CK knowledge base, they didn't analyze and interpret MITRE ATT&CK evaluations.

Other commercial security 'benchmarks' apply miscellaneous self-designed metrics in various testing environments, purely focusing on comparing EDR vendors on behalf of the customers for marketing purposes. For instance, Gartner tries to address the benchmarking challenges with its Magic Quadrant [19]. However, the methodology is not transparent and is lack of explanation. Moreover, Gartner emphasizes secondary concerns for businesses like value and viability, which don't provide insights on improving the security systems' performance. Another attempt in the industry is AV-Comparatives [6], which evaluates the anti-virus capabilities of security products. However, the evaluation methodology is not transparent like Magic Quadrant's, and the evaluations only focus on a narrow range of attack techniques.

7 DISCUSSION

An important problem we could not address in this paper is missing information, including but not limited to false positive alarm volume, response time, and raw data.

False positive alarm volume is an important indicator of manpower needed for using the EDR system [2, 36]. Low false positive volume means most alarms are true positive so that the system administrator can focus on mitigation. On the other hand, a high false alarm volume means many of the alarms are false positive alarms. Hence, the system administrator must discover true positive alarms before mitigating attacks, often leading to a needle-in-a-haystack problem. Response time indicates the time elapsed between compromises and alarm generation, which measures the real-time capabilities of the EDR systems. Low response time means the system can detect threats fast so that the system administrator can keep the loss to a minimum. Although the delayed modifier gives some information about the delayed alarm, it does not provide quantitative data on how long the delay is. Moreover, MITRE only provides the detection results from the EDR systems, not the raw data such as system logs and network events. Missing the raw data prevents new EDR systems from using the same dataset to compare performance with existing ones and limits the information available to researchers when they analyze threats with poor detection coverage. Missing the information described above hinders the analysis of EDR systems from many meaningful perspectives. We hope MITRE could include them in the future release of evaluations.

Another concern is the prospective compatibility of our interpretation framework within the context of the MITRE evaluation. While MITRE implemented notable modifications to its evaluation framework during the initial three rounds, the framework employed in the latest three rounds has demonstrated sustained consistency. Our interpretation comprehensively encompasses all elements that have maintained this consistency throughout these rounds. Therefore, we are confident that our interpretation framework will remain pertinent and enduring in the foreseeable future.

8 CONCLUSION

By leveraging MITRE's evaluation efforts and introducing our analysis method, we offer valuable insights into the current capabilities of industrial EDR systems to bridge the gap between MITRE's raw evaluation results and comprehensive interpretations. This research aids researchers, practitioners, and vendors in understanding the strengths, limitations, and areas for improvement of EDR systems, ultimately enhancing enterprise security.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant no. 2148177 and is supported in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program.

REFERENCES

[1] [n. d.]. *Endpoint Detection and Response Market Report.* https://www.mordorintelligence.com/industry-reports/endpoint-detection-and-response-market

[2] Bushra A. Alahmadi, Louise Axon, and Ivan Martinovic. 2022. 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 2783–2800. https://www.usenix.org/conference/usenixsecurity22/presentation/alahmadi

[3] Naif Saleh Almakhdhub, Abraham A. Clements, Mathias Payer, and Saurabh Bagchi. 2019. BenchIoT: A Security Benchmark for the Internet of Things. In *49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*.

[4] Abdulellah Alsaheel, Yuhong Nan, Shiqing Ma, Le Yu, Gregory Walkup, Z Berkay Celik, Xiangyu Zhang, and Dongyan Xu. 2021. ATLAS: A Sequence-based Learning Approach for Attack Investigation.. In *USENIX Security Symposium*. 3005–3022.

[5] Adel Alshamrani, Sowmya Myneni, Ankur Chowdhary, and Dijiang Huang. 2019. A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities. *IEEE Communications Surveys Tutorials* 21, 2 (2019). https://doi.org/10.1109/COMST.2019.2891891

[6] AV-Comparatives. 2023. AV-Comparatives. https://www.av-comparatives.org/.

[7] Brian Barrett. 2020. How 4 Chinese Hackers Allegedly Took Down Equifax. https://www.wired.com/story/equifax-hack-china/.

[8] Adam Bates, Dave (Jing) Tian, Kevin R.B. Butler, and Thomas Moyer. 2015. Trustworthy Whole-System Provenance for the Linux Kernel. In *24th USENIX Security Symposium (USENIX Security 15)*. USENIX Association, Washington, D.C., 319–334. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/bates

[9] William J. Broad, John Markoff, and David E. Sanger. 2011. Israeli Test on Worm Called Crucial in Iran Nuclear Delay. https://www.nytimes.com/2011/01/16/world/middleeast/16stuxnet.html.

[10] Seungoh Choi, Jeong-Han Yun, and Byung-Gil Min. 2021. Probabilistic Attack Sequence Generation and Execution Based on MITRE ATT&CK for ICS Datasets. In *Cyber Security Experimentation and Test Workshop* (Virtual, CA, USA) *(CSET '21)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/3474718.3474722

[11] CNNMoney. 2013. Target: 40 million credit cards compromised. https://money.cnn.com/2013/12/18/news/companies/target-credit-card/index.html.

[12] CompaniesMarketCap. 2021. IT Security - Largest Companies by Market Cap. https://companiesmarketcap.com/it-security/largest-companies-by-market-cap/. [Online; accessed 6-March-2023].

[13] Crowdstrike. 2021. CrowdStrike Achieves 100% Detection Coverage. https://www.crowdstrike.com/blog/crowdstrike-falcon-mitre-attack-evaluation-results-third-iteration/.

[14] CrowdStrike. 2023. *Kerberoasting - Cybersecurity 101.* https://www.crowdstrike.com/cybersecurity-101/kerberoasting/

[15] DARPA. 1998. DARPA Intrusion Detection Evaluation. https://archive.ll.mit.edu/ideval/index.html

[16] DARPA. 2023. DARPA Transparent Computing. https://www.darpa.mil/program/transparent-computing

[17] Abhishek Divekar, Meet Parekh, Vaibhav Savla, Rudra Mishra, and Mahesh Shirole. 2018. Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives. In *IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*.

[18] Center for Threat-Informed Defense. Accessed: 2023. Adversary Emulation Library. https://github.com/center-for-threat-informed-defense/adversary_emulation_library.

[19] Gartner. 2021. Magic Quadrant for Endpoint Protection Platforms. https://www.gartner.com/en/documents/4001307/magic-quadrant-for-endpoint-protection-platforms.

[20] Michael Grace, Yajin Zhou, Qiang Zhang, Shihong Zou, and Xuxian Jiang. 2012. RiskRanker: Scalable and Accurate Zero-Day Android Malware Detection. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*.

[21] Xueyuan Han, Thomas Pasquier, Adam Bates, James Mickens, and Margo Seltzer. 2020. Unicorn: Runtime Provenance-Based Detector for Advanced Persistent Threats. *Network and Distributed System Security Symposium* (2020).

[22] Gaojian Hao, Feng Li, Wei Huo, Qing Sun, Wei Wang, Xinhua Li, and Wei Zou. 2019. Constructing Benchmarks for Supporting Explainable Evaluations of Static Application Security Testing Tools. In *International Symposium on Theoretical Aspects of Software Engineering (TASE)*.

[23] Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. 2019. NoDoze: Combatting Threat Alert Fatigue with Automated Provenance Triage. In *Network and Distributed System Security Symposium*.

[24] Md Nahid Hossain, Sadegh M. Milajerdi, Junao Wang, Birhanu Eshete, Rigel Gjomemo, R. Sekar, Scott Stoller, and V.N. Venkatakrishnan. 2017. SLEUTH: Real-time Attack Scenario Reconstruction from COTS Audit Data. In *USENIX Security Symposium*.

[25] Md Nahid Hossain, Sanaz Sheikhi, and R. Sekar. 2020. Combating Dependence Explosion in Forensic Analysis Using Alternative Tag Propagation Semantics. In *IEEE Symposium on Security and Privacy (SP)*.

[26] Samuel T. King and Peter M. Chen. 2003. Backtracking Intrusions. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (Bolton Landing, NY, USA) *(SOSP '03)*. Association for Computing Machinery, New York, NY, USA, 223–236. https://doi.org/10.1145/945445.945467

[27] Zhenyuan Li, Qi Alfred Chen, Runqing Yang, Yan Chen, and Wei Ruan. 2021. Threat detection and investigation with system-level provenance graphs: a survey. *Computers & Security* 106 (2021), 102282.

[28] Shiqing Ma, Xiangyu Zhang, and Dongyan Xu. 2016. ProTracer: Towards Practical Provenance Tracing by Alternating Between Logging and Tainting. In *Network and Distributed System Security Symposium*.

[29] Angelos K. Marnerides, Petros Spachos, Periklis Chatzimisios, and Andreas U. Mauthe. 2015. Malware detection in the cloud under Ensemble Empirical Mode Decomposition. In *International Conference on Computing, Networking and Communications (ICNC)*.

[30] Naaliel Mendes, Henrique Madeira, and Joao Duraes. 2014. Security Benchmarks for Web Serving Systems. In *IEEE 25th International Symposium on Software Reliability Engineering*.

[31] Sadegh M. Milajerdi, Birhanu Eshete, Rigel Gjomemo, and Venkat Venkatakrishnan. 2019. POIROT: Aligning Attack Behavior with Kernel Audit Records for Cyber Threat Hunting. *ACM SIGSAC Conference on Computer and Communications Security* (2019).

[32] Sadegh M. Milajerdi, Rigel Gjomemo, Birhanu Eshete, R. Sekar, and V.N. Venkatakrishnan. 2019. HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows. In *IEEE Symposium on Security and Privacy (SP)*.

[33] MITRE. 2023. ATT&CK® EVALUATIONS. https://attackevals.mitre-engenuity.org/.

[34] MITRE. 2023. Matrix - Enterprise | MITRE ATT&CK®. https://attack.mitre.org/matrices/enterprise/.

[35] Alexander V. Outkin, Patricia V. Schulz, Timothy Schulz, Thomas D. Tarman, and Ali Pinar. 2023. Defender Policy Evaluation and Resource Allocation With MITRE ATT&CK Evaluations Data. *IEEE Transactions on Dependable and Secure Computing* 20, 3 (2023), 1909–1926. https://doi.org/10.1109/TDSC.2022.3165624

[36] Georgios P. Spathoulas and Sokratis K. Katsikas. 2009. Using a Fuzzy Inference System to Reduce False Positives in Intrusion Detection. In *2009 16th International Conference on Systems, Signals and Image Processing*. 1–4. https://doi.org/10.1109/IWSSIP.2009.5367701

[37] Erik van der Kouwe, Gernot Heiser, Dennis Andriesse, Herbert Bos, and Cristiano Giuffrida. 2019. SoK: Benchmarking Flaws in Systems Security. In *IEEE European Symposium on Security and Privacy (EuroS P)*.

[38] Q. Wang, Wajih Ul Hassan, Ding Li, Kangkook Jee, X. Yu, Kexuan Zou, J. Rhee, Zhengzhang Chen, Wei Cheng, Carl A. Gunter, and Haifeng Chen. 2020. You Are What You Do: Hunting Stealthy Malware via Data Provenance Analysis. In *Network and Distributed System Security Symposium*.

[39] Yulai Xie, Dan Feng, Yuchong Hu, Yan Li, Staunton Sample, and Darrell Long. 2020. Pagoda: A Hybrid Approach to Enable Efficient Real-Time Provenance Based Intrusion Detection in Big Data Environments. *IEEE Transactions on Dependable and Secure Computing* 17, 6 (2020), 1283–1296. https://doi.org/10.1109/TDSC.2018.2867595

[40] Yulai Xie, Yafeng Wu, Dan Feng, and Darrell Long. 2021. P-Gaussian: Provenance-Based Gaussian Distribution for Detecting Intrusion Behavior Variants Using High Efficient and Real Time Memory Databases. *IEEE Transactions on Dependable and Secure Computing* 18, 6 (2021), 2658–2674. https://doi.org/10.1109/TDSC.2019.2960353

[41] Chunlin Xiong, Tiantian Zhu, Weihao Dong, Linqi Ruan, Runqing Yang, Yan Chen, Yueqiang Cheng, Shuai Cheng, and Xutong Chen. 2020. CONAN: A Practical Real-time APT Detection System with High Accuracy and Efficiency. *IEEE Transactions on Dependable and Secure Computing* (Feb. 2020).

[42] Jun Zeng, Zheng Leong Chua, Yinfang Chen, Kaihang Ji, Zhenkai Liang, and Jian Mao. 2021. WATSON: Abstracting Behaviors from Audit Logs via Aggregation of Contextual Semantics.. In *NDSS*.

[43] Jun Zengy, Xiang Wang, Jiahao Liu, Yinfang Chen, Zhenkai Liang, Tat-Seng Chua, and Zheng Leong Chua. 2022. Shadewatcher: Recommendation-guided cyber threat analysis using system audit records. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 489–506.

[44] Michael Zipperle, Florian Gottwalt, Elizabeth Chang, and Tharam Dillon. 2022. Provenance-Based Intrusion Detection Systems: A Survey. *ACM Comput. Surv.* 55, 7, Article 135 (dec 2022), 36 pages. https://doi.org/10.1145/3539605

[45] Richard Zuech, Taghi M. Khoshgoftaar, Naeem Seliya, Maryam Mousaarab Najafabadi, and Clifford Kemp. 2015. A New Intrusion Detection Benchmarking System. In *Florida Artificial Intelligence Research Society Conference*.

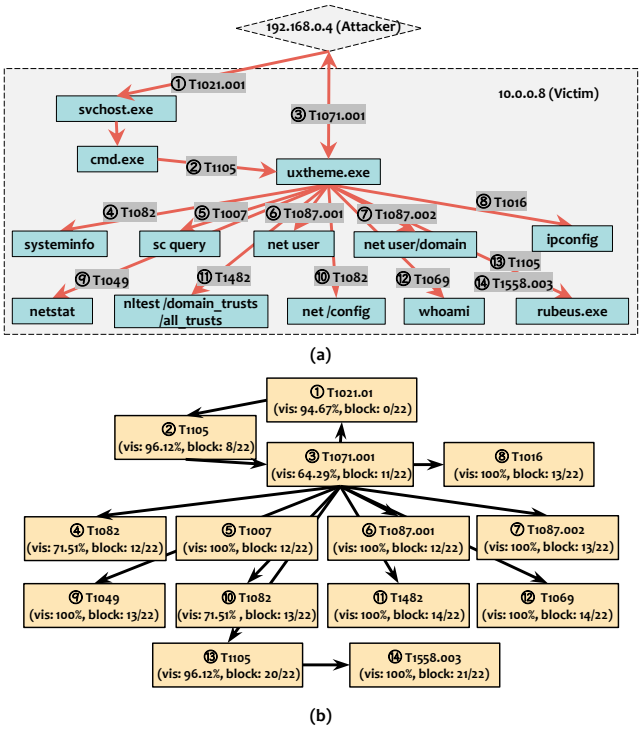


Figure 5: The actual attack graph and the causal relationship attack graph for scenario 2 in Wizard Spider+Sandworm (2022) evaluation.

A APPENDIX

A.1 Ethics

This study does not raise any ethical issues. All the datasets we used are publicly available and anonymized.

A.2 Additional graphs

Fig. 5 is another causal relationship attack graph we constructed. Fig. 6 shows the distribution of visibility and analytic scores of all EDR systems. Fig. 7 shows the same distribution of all techniques.

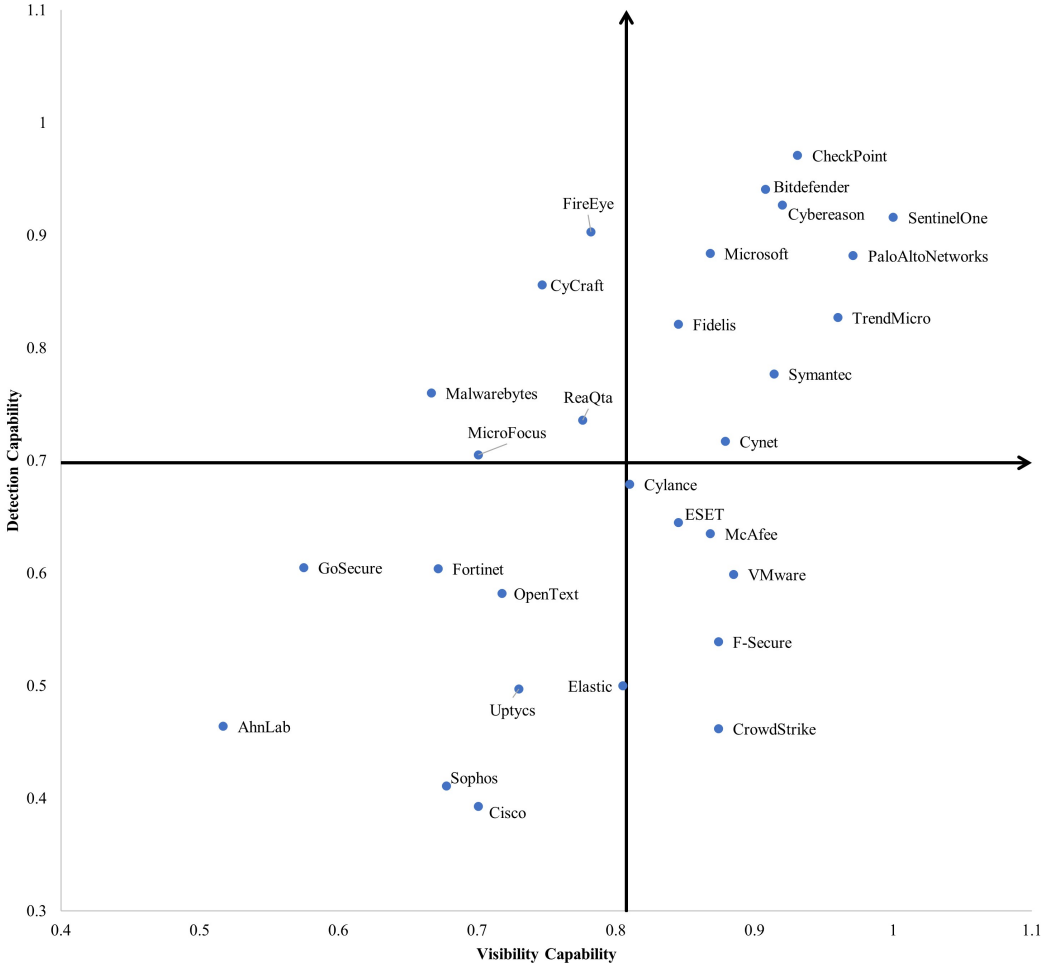


Figure 6: Analytics vs. Visibility Quadrant for EDR Systems

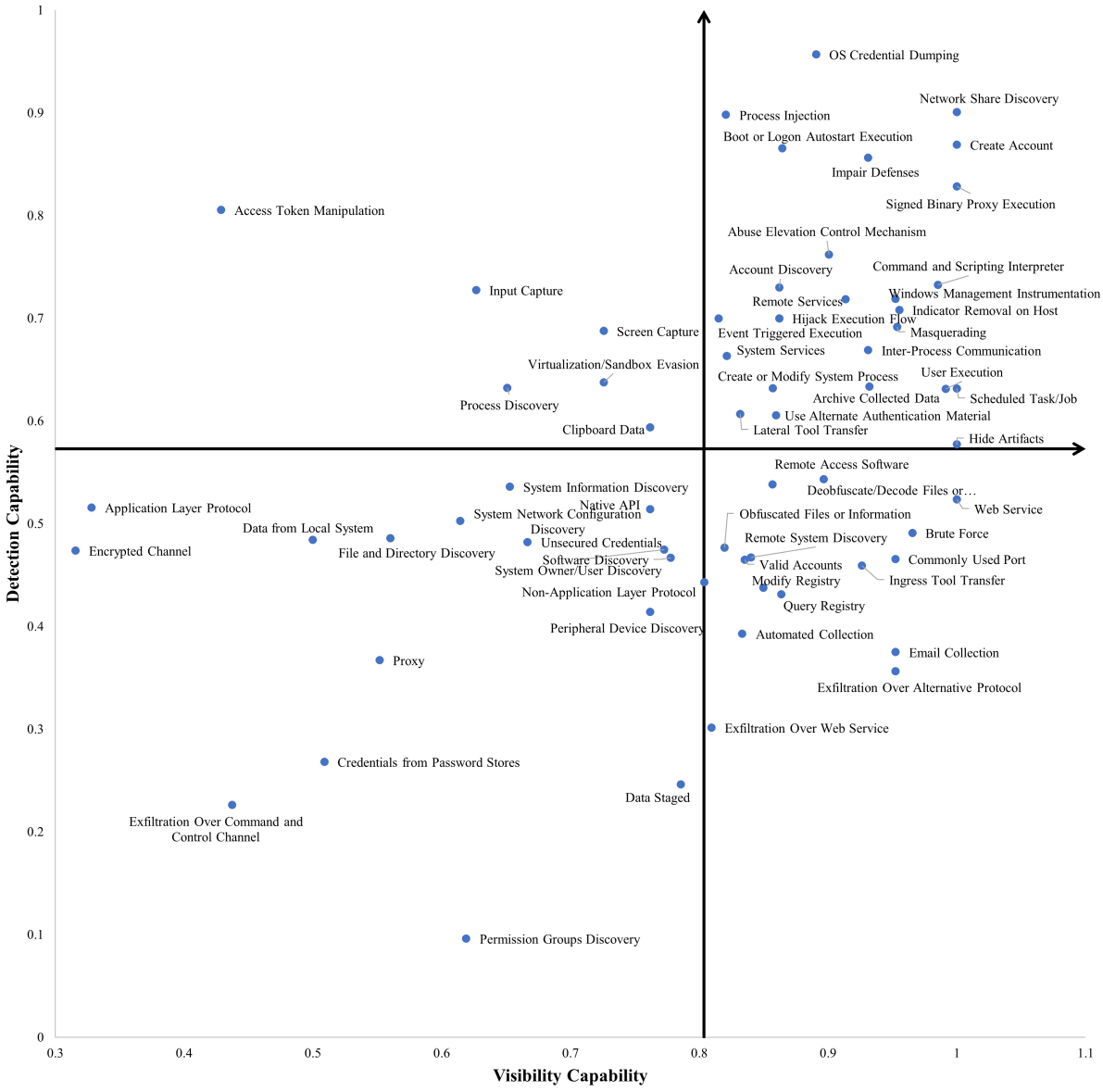


Figure 7: Analytics vs. Visibility Quadrant for Techniques