Sharing and Automation for
Privacy Preserving Attack Neutralization

(H2020 833418)

D4.4 Algorithms to recommend response and recovery actions to human operators (M15)

Published by the SAPPAN Consortium

Dissemination Level: Public

H2020-SU-ICT-2018-2020 - Cybersecurity

Document control page

Document file: Deliverable
Document version: 1.0
Document owner: Alexey Kirichenko (F-Secure)

Work package: WP4
Task: T4.3
Deliverable type: Other
Delivery month: M15 (July 2020)
Document status: ☒ approved by the document owner for internal review
☒ approved for submission to the EC

Document History:

Version	Author(s)	Date	Summary of changes made
0.1	Andrew Patel, Paolo Palumbo, Alexey Kirichenko (F-Secure)	2020-07-21	Preliminary document sent out for comments by all partners
0.2	Alexey Kirichenko	2020-07-23	Integrated minor revisions by Avikarsha Mandal and Matti Aksela (internal FSC review)
0.3	Alexey Kirichenko	2020-07-25	Integrated minor revisions by Martin Zadnik
1.0	Avikarsha Mandal	2020-07-31	Integrated editorial comments from Mischa Obrecht, ready for submission

Internal review history:

Reviewed by	Date	Summary of comments
Avikarsha Mandal	2020-07-23	Looks fine, minor Editorial Comments: • Change the Dissemination Level from *confidential* to *public* • Change the Abstract to "Executive Summary" • Page 12, 3rd paragraph: i) how to strategic select → i) how to strategically select
Martin Zadnik	2020-07-24	
Sarka Pekarova	2020-07-24	
Mischa Obrecht	2020-07-28	Nice read! Very well structured and easily comprehensible. Two minor comments: • Section 2.1, "Technical security Controls, an Overview": There is no mention of any device that regulates network traffic, such as Firewall, IDS, IPS, etc. It might be worth considering a fourth class of security control. • Section 5, "Acknowledgments": There's a typo, please call us "Dreamlab Technologies AG" instead of "... AGW".

Authors

Paul Blomstedt, Senior Data Scientist at Artificial Intelligence Center of Excellence, F-Secure Corporation (Paul.Blomstedt@F-Secure.com)
Jouni Kallunki, Senior Manager at Artificial Intelligence Center of Excellence, F-Secure Corporation (jouni.kallunki@f-secure.com)
David Karpuk, Senior Data Scientist at Artificial Intelligence Center of Excellence, F-Secure Corporation (David.Karpuk@F-Secure.com)
Alexey Kirichenko, Research Collaboration Manager, F-Secure Corporation (alexey.kirichenko@f- secure.com)
Dmitriy Komashinskiy, Lead Researcher at Artificial Intelligence Center of Excellence, F-Secure Corporation (dmitriy.komashinskiy@f-secure.com)
Andrew Patel, Researcher at Artificial Intelligence Center of Excellence, F-Secure Corporation (andrew.patel@f-secure.com)
Paolo Palumbo, Director, Protection Strategy, F-Secure Corporation (paolo.palumbo@f-secure.com)
Tomas Jirsik, RNDr, Ph.D., CSIRT-MU, Masaryk University (jirsik@ics.muni.cz)
Martin Zadnik, Ing., Ph.D., Project Manager and researcher at Traffic Monitoring and Configuration Department, CESNET (zadnik@cesnet.cz)

1 Executive Summary

Defending against cyber-attacks remains a challenging task, especially given the lack of experts in the cybersecurity field. Organizations are attempting to solve this problem by deploying tools that enable less experienced security analysts to perform at a higher level of expertise. When working with incident response systems, analysts often deal with a large number of false alerts. False alerts can outnumber true attack detections by a factor of 10 or even more. When an analyst spends most of their shift separating false positives from actionable incidents, fatigue can set in, and real incidents can go unnoticed. One area of particular interest to cybersecurity tool vendors, and the main focus of this report, is the automation of incident response recommendation mechanisms that are able to filter out many of these false positives. It is worth noting that an automated system that can recognize true alerts (that require response actions) is the first step towards a truly automated response system. Understanding the type and severity of a security incident that triggered an alert can then be used to choose appropriate response actions, which can be suggested to security personnel or, in certain cases, even carried out automatically. This report is divided into two main sections. In the first section, we examine current literature on this subject from the cybersecurity industry, academia, and H2020 projects. In the second section we describe three mechanisms that are being developed within SAPPAN at F-Secure to provide response recommendations to security analysts and address the false positive problem.

Table of Contents

2 Introduction

Fueled by several factors, defending against cyber-attacks continues to be a challenging proposition. The level of threat posed by adversaries has increased both directly and indirectly. At the top end, threat actors are showing a greater level of sophistication in the tactics they employ. However, the wide availability of tools, documentation, tutorials, communities, and code snippets has made the process of performing attacks much easier. As such, at the low end, unskilled adversaries are using these tools to great effect, and this is something worth worrying about. Essentially, the return on investment for performing attacks has increased, resulting in an uptick in breaches, regardless of vertical or organization [1]. This situation is exasperated by the fact that skilled cybersecurity specialists are still a rare commodity. Companies have trouble retaining experts due to poaching, and it is difficult to train junior talent when their few seniors are always busy responding to an onslaught of incidents.

On the defense side, technical security controls – the tools that an organization can put in place to protect itself – are diverse both in terms of their capabilities and the format in which they provide information. As such, these tools have a steep learning curve, and transitioning between different vendors' tools requires considerable effort. Many organizations are still struggling to define suitable security processes and to deploy even basic security controls.

Experts are still in high demand because much of the work that goes into detecting and responding to security incidents is still manual. By automating some of this work, companies can effectively manage their security with fewer experts, in a more sustainable fashion. Easier, better tools allow companies to train new hires quicker and be resilient to the loss of experienced personnel. Automation provides other benefits to a company's security operations such as streamlining of workflows, operational visibility, better tracking of key performance indicators, and an improvement of mean time to repair in the context of security incidents. However, the automation of activities related to cybersecurity remains exclusive to a select few mature organizations. Very few companies perform cyber threat intelligence or threat hunting. In some cases, organizations are even reluctant to deploy automation out of fear that it may dull the skill of their human experts.

Automation introduces risks of incorrect action, especially in cases when additional contextual information, available to experts, is not considered during the design of automation processes or in cases when attackers succeed in tricking the automation. Experts are also concerned about non-deterministic and non-verifiable algorithms which may fail when they encounter corner cases. Although ongoing efforts, such as development of explainable AI algorithms, aim to address the mentioned concerns, it is often safer to start with automation mechanisms with low severity of incorrect decisions and providing more control to the expert.

2.1 Technical security controls, an overview

Technical security controls fall into three rough categories:

- Endpoint Detection and Response (EDR)
- Security Orchestration, Automation, and Response (SOARs)
- Security Information and Event Management (SIEMs)

Endpoint detection and response platforms deploy agents to each protected endpoint to gather data. This data is analyzed to reveal potential cyber threats and issues. EDR solutions protect against hacking by continually monitoring each endpoint and storing all data in a secure location where it cannot be tampered with. When an incident is detected, the end-user is immediately

prompted with a list of preventative actions. Although different EDR platforms boast different capabilities, they share a common set of functionalities that include monitoring endpoints both online and offline, real-time response to discovered threats, improved visibility and transparency of user data, detection of malware injection events, allow/deny lists and integration with other security technologies [2][3].

Gartner defines security orchestration, automation and response (SOAR) as technologies that enable organizations to take inputs from a variety of sources (mostly from security information and event management (SIEM) systems) and apply workflows aligned to processes and procedures. These can be orchestrated via integrations with other technologies and automated to achieve a desired outcome and greater visibility. Additional capabilities include case and incident management features; the ability to manage threat intelligence, dashboards and reporting; and analytics that can be applied across various functions. SOAR tools significantly enhance security operations activities like threat detection and response by providing machine-powered assistance to human analysts to improve the efficiency and consistency of people and processes.[4]

Security information and event management (SIEM) systems are designed to support threat detection, compliance, and security incident management through the collection and analysis of security events and other contextual data sources. They do this by aggregating logs from systems in an organization, analyzing those logs, and presenting dashboard visualizations to security analysts. SIEM core capabilities, across different vendor offerings, include log event collection and management, the ability to analyze log events and other data across disparate sources, and operational capabilities such as incident management, dashboards, and reporting [5].

Although we have grouped technical security control solutions into three categories, the capabilities of different vendor offerings are starting to overlap. Some SIEM solutions now offer capabilities found in the traditional SOAR scope, and increasingly EDR offerings have started adding functionality that could be considered part of the SIEM and SOAR domain. By and large, technical security control solutions have recently introduced automation to address the challenges mentioned earlier. This automation tends towards the following.

- Mechanisms to suggest actions based on specific triggers. For instance, if a network connection to a malicious domain is detected, isolate the host that made the connection. If an unknown file is executed, submit the file for automated sandbox detonation. These mechanisms are often implemented as pre-defined rules and logic, some of which is shipped with the product. Users of the solution can also add their own rules.
- Mechanisms that execute actions from common workflows and playbooks. For instance, when a phishing attempt is detected, reset the user's password and contact the user. These workflows can be set to execute automatically, or under full or partial human supervision. When the steps in a workflow involve information gathering, automation will retrieve that data, sparing the analyst from that manual task. These automation mechanisms are common in SOAR solutions.
- Mechanisms that provide contextual recommendations, such as how to respond to a threat, how to proceed with an investigation, or how to perform threat hunting. Products with these capabilities tend to have their origin in the more general area of IT automation and have developed this new functionality in response to cybersecurity trends. A good example is ServiceNow.

2.2 Response automation functionality by vendor

In this section we examine detection and response automation capabilities provided by specific cybersecurity vendors' solutions. Unfortunately, most cybersecurity vendors do not elaborate on their specific methodologies and occasionally exaggerate on the capabilities of their solutions, especially

with regards to their use of artificial intelligence. As such, this section contains what we considered to be relevant snippets and quotes from marketing literature we were able to find.

Symantec Endpoint Security Complete

Symantec's solution offers EDR capabilities promising "prevention across the whole kill chain". According to the product brief, "AI-guided security management more accurately updates policies, with fewer misconfigurations to improve your security hygiene" and "autonomous security management continuously learns from administrator and user behaviors to improve threat assessments, tune responses, and strengthen your overall security posture"[6]. Additionally, the solution includes "built-in playbooks that encapsulate the best practices of skilled threat hunters and anomalous behavior detection". Finally, the solution supports automatic submission of identified suspicious files to sandboxing for complete malware analysis including exposing malware that is VM-aware [6]

McAfee MVISION

McAfee's MVISION [8] is also an EDR product. Among the many features offered by the solution, the company highlights that the product offers "AI guided investigations" and such technology that "allows tier 1 analysts to operate as seasoned veterans" [9]. In more detail "MVISION EDR automatically gathers, summarizes, and visualizes evidence from multiple sources and iterates as the investigation evolves" [10]. "The AI-powered investigation engine gathers and processes artifacts and complex event sequences ... to help make sense of alerts. MVISION EDR compares evidence against known normal activity for each organization and threat intelligence sources to improve local relevancy and reduce false positives triggered against normal activity [10]

TrendMicro XDR

TrendMicro's XDR-powered solution [11] offers detection and hunting capabilities, including YARA integration, and shows clearly the company's cybersecurity heritage. TrendMicro's solutions offer the capability of providing an 'automated root cause analysis', that helps analysts scan other relevant assets for signs of similar infections [12]. This feature seems to be particularly effective for addressing email-based threats. The company's relevant products also offer some level of automatic remediation for specific threats, such as specific strains of ransomware, and often allow the end users to customize automatic actions that are taken in reaction to specific alerts or detections [13].

Sophos Intercept X

Sophos Intercept X [14] falls into the EDR category and offers a series of capabilities that facilitate the work of a SOC analyst. For example, Intercept X products offer a 'Threat Indicators' section in their user portal, that shows suspect portable executables, ranked by suspiciousness and prevalence. This capability enables analysts to direct their attention to items that require it the most [15]. The products also feature an automated root cause analysis technology [16] which automatically collects information about certain kinds of alert and presents it to the case handler. According to Sophos, this technology is designed to answer the "what, where, when and how" questions. Finally, some of the company's higher-tier products feature 'automated threat hunting technology' [17], but details about this technology are scarce; it is likely that these capabilities were the result of the acquisition of DarkBytes [18]

ATAR Labs ATAR

ATAR Labs' ATAR [19] is a solution that helps manage SOC activities by offering three main capabilities – playbooks and automation, incident management, and SOC analytics. ATAR provides comprehensive automation and tight SIEM integrations. ATAR also has capabilities

to monitor key performance indicators via customizable dashboards. Belonging to the SOAR category, ATAR promises to provide value across three different dimensions: (i) automation of repetitive activities, (ii) improvement of analyst efficiency and (iii) increased ability to measure performance via tracking of KPIs. When it comes to automation [20], the solution allows analysts to automate frequent scenarios, and provides a wide array of integrations that can collect and pre-contextualize the alerts that are then passed on to the analyst. Their literature states "by using ATAR, SOC teams can pass all repetitive activities to platform and whenever an incident occurs ATAR will handle it without human interaction. ATAR also allows to bring the incident up to a certain point that human analyst can take over from that point and continue to work on incident. When a new hire arrives at the SOC, (s)he is given playbooks describing what to do in the occurency of a particular type of incident." (sic)

Ayehu NG

“Founded in 2007, the Ayehu NG platform is a web-based IT automation and orchestration solution for security and IT operations. Its key features are playbook scheduling, enabling selective alerts to support remote control of incidents, audit trail generation, rollback of changes to workflows and role-based access to workflows in order to maintain access, segregation for both teams (IT and security). Also, Ayehu NG uses machine learning to suggest playbooks and creation of rules. In addition, Ayehu NG bridges the gap between IT and security operations (network operations center [NOC] and SOC), streamlining automated workflow processes and tasks, and resolving IT and security alerts and incidents to improve SLAs” [21]. The product descriptions highlight features meant to leverage automation such as "machine learning driven decision support", which does 'provide decision support via suggestions to optimize your workflows and dynamically create rule-based recommendations and insights', by 'leveraging proprietary, sophisticated machine learning algorithms'.

EclecticIQ

EclecticIQ is a company that focuses on Cyber Threat Intelligence [22]. EclecticIQ provides benefits across role types thanks to its ability to contextualize operations via threat intelligence and by providing automation. According to further material [23], the platform "provides a core set of workflows within a single collaborative workspace. Using these workflows, analysts within Security Operations Centers (SOCs), Computer Emergency Response Teams (CERTs), Fusion Centers, Intelligence Teams and Threat Hunting Teams can quickly discern actionable and relevant intelligence, collaborate with other analysts, update enterprise security controls and share information with external communities". Triaging and prioritization are achieved via "policy-based alerts based on advanced search logic and network graph correlation matrices and by qualifying threats based on proximity, confidence, threat level or other factors fully customizable to your own workflow and taxonomy."

Phantom

Phantom is a SOAR solution built on top of the analytic capabilities of Splunk [24]. Phantom Playbooks [25] execute sets of actions across security infrastructure, allowing analysts to 'automate actions at security speed'. Phantom comes with 100+ pre-built playbooks that can be edited and designed using a visual tool that does not require programming. Additionally, the solution offers a feature called ‘Mission Guidance’ [26], which is, in the company’s own words, an “intelligent assistant that supports security operations analysts. Phantom Mission Guidance offers suggestions to help investigate, contain, eradicate and recover from a security event. It works by mapping security event data to your currently configured SOC tools and playbooks. Phantom Mission Guidance recommendations help educate newer analysts on steps to take and validate the choices of more experienced analysts”.

Panda

Panda's approach to automating threat detection, investigation, and response is heavily biased towards fully automating operations. In their higher tier products [27], they deploy a technology known as Adaptive Defense 360 that promises (i) automatic and transparent remediation and (ii) actionable insights into attackers and their activity, speeding up forensic investigation. The company is very clear in its intent to eliminate end user interaction by providing a fully automatic solution. Further automation is supported in an add-on product, the Advanced Reporting Tool for Adaptive Defense 360. This additional module can "automatically generate security intelligence and provide tools that allow organizations to pinpoint attacks and unusual behaviors, regardless of their origin, as well as detecting internal misuse of the corporate network and systems" by "automating the storage and correlation of information generated by the execution of processes and their context, extracted from endpoints by Panda Adaptive Defense 360" [28].

VMWare Carbon Black

VMWare's Carbon Black-related products [29][30] are based on a single endpoint agent and on a unified technology stack. The products seem to provide responders and investigators with a prioritized list of alerts to look at, hence optimizing "Mean-Time-To-Resolution". Their EDR products [31] are designed to allow the management of vast fleets of endpoints, providing SOC operators with the visibility that they need. This includes the capability of conducting hunts on the sensor estate. Their dashboards show signs of scoring and similar mechanism for ranking artifacts and incidents, so that they can be visually presented in a way to aid the operator.

Crowdstrike

Crowdstrike's offering is organized as a set of products and services that build upon each other's capabilities in a tiered fashion. Crowdstrike's threat intelligence component, Falcon X [32] offers "automated Investigations", that "bring endpoint protection to the next level by combining malware sandbox analysis, malware search and threat intelligence in a single solution". According to the associated material, the solution provides automatic investigation aid, for example by automatically submitting suspicious files to a sandbox for detonation, or by providing automated threat actor attribution, or presenting lists of related artifacts as a matter of providing context and aiding investigations. The company's additional EDR component, Falcon Insight [33] features "smart prioritization" which "automates triage and shows you what deserves attention first". Smart prioritization is done through Crowdscore [34][35], which is a measure of the severity of an incident and can be used to take a priority-based approach at handling cases. Crowdscore can also be applied to organizations as a whole.

LogRhythm

LogRhythm [36] provides a 'next-generation SIEM platform', which provides "intuitive, high-performance analytics and a seamless incident response workflow". The product provides a set of pre-built playbooks that an analyst can select, which in turn contain sequences of actions that can/should be taken as part of the remediation effort. LogRhythm's risk score has been patented [37][38]. Based on available material, the solution can either pre-suggest case playbooks [39] or allow them to be manually selected by the analyst. Response actions can also be pre-suggested (i.e. "approval driven") or manually triggered.

Cylance

Cylance OPTICS is Cylance's EDR offering [40]. According to the solution brief [41], Cylance OPTICS "deploys trained threat behavior models directly on the endpoint. This em-

powers protected devices to function as self-contained security operations centers (SOC), independent of cloud connectivity. Cylance OPTICS includes a configurable context analysis engine (CAE) that monitors endpoint events in near real time".

Arbor

Arbor's Sightline with Sentinel "combines the technologies of NETSCOUT and Arbor to deliver smarter traffic visibility and threat detection, as well as an automated, fully integrated DDoS defense" [42]. According to the company's material [43], Arbor's Netscout technology can automatically detect and mitigate DDOS attacks. It acts by orchestrating mitigating actions across the network and beyond. It is an 'always on technology that continuously absorb layer 7 data and ATLAS intelligence'. Netscout relies on a community of large network providers sharing data and promises an inter-organization response and defense against DDOS. Arbor's Threat Mitigation System [44] is the component that offers DDOS protection. It can automatically detect DDOS by leveraging "statistical anomaly detection, protocol anomaly detection, fingerprint matching and profiled anomaly detection." Their literature goes on to state "our solution continually learns and adapts in real-time, alerting operators to attacks, as well as to unusual changes in demand and service levels. Arbor TMS can isolate and remove the attack traffic, without affecting other users, in as fast as a few seconds. Methods include identifying and black-listing malicious hosts, IP location-based mitigation, protocol anomaly-based filtering, malformed packet removal and rate limiting (to gracefully manage non-malicious demand spikes). Mitigations can be automated or operator-initiated and countermeasures can be combined to address blended attacks".

RESPOND ANALYST

RESPOND ANALYST [45] is a solution that aims at automating tier 1 analysts, building up cases and escalating to more senior tier analysts when needed. According to the company's documentation [46], "using patented techniques and probabilistic mathematics, the Respond Analyst monitors security event streams and automates expert human analysis of security alerts, accurately culling false positives and escalating actionable, prioritized and well-articulated incidents". According to the company's material [47] the solution's capabilities in terms of automation include investigating threats, scoping and building cases and prioritization and escalation when needed. The solution's performance can be further improved and fine-tuned through interaction with operators and other users.

ServiceNow

ServiceNow Security Operations is ServiceNow's SOAR solution built upon the Now platform [48]. The Now platform [49] is meant to 'quickly digitize workflows and run them at scale'. Their literature goes on to state "the platform is designed from the ground up with AI and predictive capabilities and we believe that these capabilities are also available for the security extension." The operation-supporting AI capabilities that are specifically mentioned in the product's documentation [50] are (i) major incident detection, (ii) action and content recommendations, (iii) categorization, routing, and prioritization and (iv) cluster analysis.

IBM

IBM offers security automation via a product called QRadar Advisor with Watson [51]. The solution "empower(s) security analysts to drive consistent, context-rich investigations to reduce dwell times and increase analyst efficiency". This product is heavily based on machine learning due to its integration with Watson. According to the IBM, "it automates routine SOC tasks, finds commonalities across investigations and provides actionable feedback to analysts, freeing them up to focus on more important elements of the investigation and increase analyst efficiency". The solution brief [52] notes that the solution provides an AI that auto-

matically finds commonalities across incidents using cognitive reasoning and provides actionable feedback with context. Additionally, the solution implements 'Easy Incident Scoring' to provide analysts with a quicker and more decisive escalation process. For those customers that choose IBM's higher tier products, the company offers Resilient [53], which is the company's SOAR offering. Resilient includes concept of Dynamic Playbooks, which allow analysts to execute playbooks in response to detection triggers.

Elastic

Elastic's security platform is an SIEM offering from the makers of the popular ELK security stack [54]. The solution promises to "easily onboard diverse data to eliminate blind spots. Surface threats with prebuilt anomaly detection jobs and detection rules. Accelerate response with a powerful investigation UI and embedded case management. All from a single UI in Kibana". Relevant features include the ability to surface anomalies through machine learning and automate detections in a way that is aligned with the MITRE ATT&CK framework.

DarkTrace

DarkTrace's products are heavily focused on automation of detection and response. DarkTrace's The Enterprise Immune System is described as a "self-learning cyber AI technology that detects novel attacks and insider threats at an early stage" [55]. The company's related technical documentation [56] claims that "Darktrace's cyber AI platform has evolved to deliver surgical automation that fights back at machine speed, taking proportionate action to contain in-progress threats before they have time to escalate into a crisis".

Palo Alto

Palo Alto's offers two relevant products that belong to the Cortex family [57], Xdr [58] and Xsoar [59]. It is through these technologies that the company implements detection and automated response solutions. Xsoar, a technology that was originally developed by Demisto, allows end customers to easily model and automate their workflows and playbooks. Based on the same material, [60] Demisto's technology incorporates a unique approach to end user-analyst interaction, in the form of the Xsoar DBot [61]. In the words of the company, "Demisto Enterprise also leverages the power of machine learning through DBot to act as a force multiplier and prime SOCs for the future. ML-supported suggestions are present in incident ticketing, task-analyst matching, response actions, analyst ownership, and related incidents. Machine learning cuts across all three pillars of case management, intelligent automation and orchestration, and interactive investigation. As both DBot and analysts grow smarter with each incident, the marginal time to investigate and respond to threats decreases".

In summary, it appears that many vendors have implemented useful point-and-click functionalities for common tasks, such as data collection and investigative work. In addition, some basic machine learning-based functionalities, such as clustering, recommendations, and classification models are likely built-in to some vendors' analyst user interfaces.

2.3 Academic literature

Some academic researchers have taken an interest in examining challenges in the incident response field. However, the number of publications in this area remains fairly low. While some research teams have attempted to create action recommendation models in the cybersecurity space, very little research has been performed in this area as compared to other areas in the machine learning space. This section examines the academic research in this area that we could find. We've broken it down into three subcategories.

Discussion of incident handling processes

"Optimal Countermeasures Selection Against Cyber Attacks: A Comprehensive Survey on Reaction Frameworks" [62] is an article that is based on a survey. The authors define the concept of a "countermeasure", provide an overview of attack modeling techniques and then discuss various standardization efforts for security automation. The work closes with a description of the research challenges, which the authors identify as (i) scalability of the systems of automated countermeasures, (ii) countermeasure knowledge management, (iii) missing standardized representation for countermeasures and in (iv) metrics for scoring the countermeasures.

In "Informing Hybrid System Design in Cybersecurity Incident Response" [63], the authors present insights originating from qualitative research with analysts who currently perform incident response work. The paper discusses research approaches in addressing issues in cybersecurity incident response, more specifically human-centered approaches, algorithmic and computational approaches and contextual inquiries. The paper then moves on to cover the topic of automation with analysts and highlights that opportunities for automation require stakeholders and need identification prior to development, and that they should consider maintenance workload per automated task in cost-benefit analyses. The disadvantages of automating tasks should be carefully evaluated, as they can increase task complexity and overall workload, as well as decrease entry-level analyst opportunities for problem-solving. Finally, the paper notes that automation has some clear advantages in helping decrease individual and organization workload with respect to incident response, and states that some opportunities have been clearly identified based on current perceptions of those advantages.

In "Cognitive Security for Incident Management Process" [64], the authors provide a literature review regarding processes for handling security incidents and identifying standards or guidelines published by international organizations.

In "Review of Human Decision-making during Incident Analysis" [65], the authors provide an overview of standards used to investigate incidents by incident responders. They identify relevant organizations that are contributing to the definition of (or outright providing) the standards, including standard reporting formats used in cybersecurity information exchange. The paper closes with valuable analysis of gaps in advice for making decisions, that the authors identify as (i) how to strategically select tactics (which analysis heuristic or technical tool should be employed in a particular situation and why), (ii) when an investigator is justified in generalizing (making a stronger or broader claim from singular piece of evidence) and (iii) what information to report and how to communicate it a convincing enough manner.

Incident-related models

In "Automate incident management by decision-making model" [66], the authors construct an automatic decision-making model based on data mining. When receiving an incident request, the model can identify the possible failing continuous integration systems based on historical data, predict the incident classification, and retrieve relevant information from a knowledge base of incidents.

In "Automated Event Prioritization for Security Operation Center using Deep Learning"[67], the authors present a new approach for SOC event classification whereby they identify a set of features using graphical analysis and then train a deep neural network model to classify those events.

In “Towards Predicting Cyber Attacks Using Information Exchange and Data Mining” [68], the authors present an empirical evaluation of an approach to predict attacker's activities based on information exchange and data mining. They then use sequential rule mining to identify common attack patterns and derive rules for predicting attacks. Their findings show that most of the rules display stable values of support and confidence and, thus, can be used to predict cyber-attacks in consecutive days, after mining, without the need to actualize the rules every day.

In “AIDA Framework: Real-Time Correlation and Prediction of Intrusion Detection” [69], the authors present AIDA, an analytical framework for processing intrusion detection alerts with a focus on alert correlation and predictive analytics. The framework contains components that filter, aggregate, and correlate the alerts, and predict future security events using predictive rules distilled from historical records.

Network-related research

In “Network entity characterization and attack prediction” [70], the authors propose a system that is intended for characterizing network entities and the likelihood that they will behave maliciously in the future. The system, namely Network Entity Reputation Database System (NERDS), considers all available information regarding a network entity to calculate the probability that it will act maliciously. Their experimental results show that it is indeed possible to precisely estimate the probability of future attacks from each entity using information about its previous malicious behavior and other characteristics. Ranking entities with this method has practical applications in alert prioritization, assembly of highly effective deny lists, and other use cases.

In “NERD: Network Entity Reputation Database” [71], the authors present an open database of known malicious entities on the internet called Network Entity Reputation Database. It gathers alerts from many diverse security monitoring tools and other sources and keeps detailed information about all network entities (IP addresses, ASNs, domain names, etc.) which have been reported as malicious. It also adds other related data from a multitude of sources, like whois registries, deny lists or geolocation databases. The authors then describe the data model, system architecture and technologies used, as well as some statistics from a pilot deployment of the system.

2.4 Horizon 2020 projects

The European Union is funding several cybersecurity-themed projects under the umbrella of Horizon 2020. Horizon 2020 is the European Union’s eighth framework programme for funding research, technological development, and innovation and is officially named "Framework Programme for Research and Innovation". The programme is implemented by the European Commission – the executive body of the European Union. Projects are directed by various offices including the directorate general for research and innovation, the directorate general for communications networks, content and technology, the Research Executive Agency (REA), the Executive Agency for SMEs (EASME), and the ERC Executive Agency (ERCEA). The framework programme's objective is to complete the European Research Area (ERA) by coordinating national research policies and pooling research funding in order to avoid duplication. Horizon 2020 itself is seen as a policy instrument to implement other high-level policy initiatives of the European Union, such as Europe 2020 and Innovation Union. The programme runs from 2014–2020 and provides an estimated €80 billion of funding [7].

Several projects in the Horizon 2020 programme are aimed at addressing the challenge of recommending actions to security analysts and developing automation in this operational domain. This section describes those projects.

SOCCRATES

SOCCRATES [72] intends to develop and demonstrate a security platform for Security Operation Centres (SOCs) and Computer Security Incident Response Teams (CSIRTs). This platform will be able to detect cyber-threats and prevent cyber-attacks, increasing the resilience of European organisations. The platform will be deployed in two pilot cases with complex and diverse ICT structures. The final aim is to offer the SOCCRATES platform to the market. The project's deliverables include automated and partially automated systems for cybersecurity operators – "SOCCRATES will develop and implement a new security platform for Security Operation Centres (SOCs) and Computer Security Incident Response Teams (CSIRTs), that will significantly improve an organisation's capability to quickly and effectively detect and respond to new cyber threats and ongoing attacks. The SOCCRATES Platform consists of an orchestrating function and a set of innovative components for automated infrastructure modelling, attack detection, cyber threat intelligence utilization, threat trend prediction, and automated analysis using attack defence graphs and business impact modelling to aid human analysis and decision making on response actions, and enable the execution of defensive actions at machine-speed". The main objective of SOCCRATES [73] is to develop and implement a security automation and decision support platform that enhances the effectiveness of SOC and CSIRT operations.

CyberSane

CyberSANE [74][75] will enhance the security and resilience of critical information infrastructure (CII) by providing a dynamic collaborative warning and response system. This will support and guide security officers to recognize, identify, dynamically analyze, forecast, treat and respond to advanced persistent threats and handle their daily cyber incidents utilizing and combining both structured data and unstructured data coming from social networks and the dark web. The chief objectives of this project include the design of forecasting procedures and models to assist CII operators and security experts. The project also aims to develop correlation techniques for optimization of automatic analysis of huge quantities of events, information and evidence combining both structure and unstructured data in a privacy-aware manner for malicious action identification in cyber assets such as abnormal behaviour. The project's activities will mostly involve work in the area of automated or semi-automated workflows in the domain of cybersecurity.

SPARTA

SPARTA – Strategic Programs for Advanced Research and Technology in Europe [76] aims to bring together a unique set of actors at the intersection of scientific excellence, technological innovation, and societal sciences in cybersecurity. Strongly guided by concrete and risky challenges, it will setup unique collaboration means, leading the way in building transformative capabilities and forming world-leading expertise centers. Through innovative governance, ambitious demonstration cases, and active community engagement, SPARTA aims at rethinking the way cybersecurity research is performed in Europe across domains and expertise, from foundations to applications, in academia and industry. Among the many groundbreaking deliverables that will be made available during the lifetime of this project, is T-SHARK, an "advanced SIEM and Cyber threat prevention specialized contributor for supporting cyber situational awareness capabilities" [77]. T-SHARK will include predictive threat intelligence and artificial intelligence techniques developed to analyze information monitored by heterogeneous data sources (e.g. NOCs, SOCs, SIEMs, and IDS/IPSs).

ReAct

With the belief [78] that the core of problem with software vulnerabilities is the fact that cyber attackers are almost always one step ahead of cybersecurity researchers and practitioners, the ReAct Horizon 2020 project aims to fight software exploitation and mitigate advanced cybersecurity threats in a timely fashion. This project aims to develop mechanisms that allow organizations to temporarily secure systems, via software instrumentation, as soon as they are made aware of a new vulnerability, until an official patch is published [79].

2.5 Discussion

There is a clear need to add intelligent automation to the incident response process, and this can be seen throughout the cybersecurity industry. The fact that the European Union is funding several major projects in this area attests to this need even further. The lack of academic research in this area can be attributed to the fact that realistic data is required in order to properly conduct this research and, by and large, only cybersecurity vendors have access to sufficient volumes of high-quality data relevant to this research. Horizon 2020 projects intend to solve this problem by putting cybersecurity vendors in touch with academic researchers.

Indeed, as part of the SAPPAN project (Task 4.3), F-Secure and other partners are conducting research focused on utilizing data analytics to assist security personnel in incident response activities. The next section describes three mechanisms in development at F-Secure that are aimed at finding similarities between security incidents and reducing false positives.

3 Building blocks for response recommendation and false positive suppression

In the context of augmenting human incident response work, we have conducted research using processes based on, and data gathered from an EDR-style solution. In this solution back end systems receive a stream of events from protected computers, and then generate *alerts* when malicious or anomalous behavior is detected. When a new *alert* arrives, a security analyst assesses the relevance of the alert – its type, severity, and potential risks, and then follows-up accordingly (logs the incident in a ticketing system, and performs actions based on what happened). One of the main problems in intrusion detection systems are false alerts, or false positives which cause extra burden to security analysts. False alerts can outnumber true attack detections by a factor of 10 or even more. When an analyst spends most of their shift separating false positives from actionable incidents, fatigue can set in, and real incidents can go unnoticed. Therefore, it is worth exploring approaches that can lower the number of false positives an analyst must deal with. We propose that machine learning can be used to build mechanisms capable of automatically recognizing and marking false positives as such, allowing security analysts to safely ignore them.

It is worth noting that an automated system that can recognize true alerts (that require response actions) is the first step towards a truly automated response system. Understanding the type and severity of a security incident that triggered an alert can then be used to choose appropriate response actions, which can be suggested to security personnel or, in certain cases, even carried out automatically.

Information about earlier alerts and how they were handled can be very helpful for the analyst, especially when a new arriving alert is similar to one that was handled in the past. With this information it may be possible to develop models capable of classifying alerts based on human-labeled data. The research described in this document focuses on approaches to evaluation of the similarity

of security alerts in order to support incident response analysts' decision-making processes, especially in the area of false positive reduction. While we focused on alerts and associated contextual information produced by specific security monitoring systems, in order to be able to experiment with actual data and validate our techniques, we believe that the described approaches are methodologically relevant for other cyberattack detection systems.

An ideal and conceptually straightforward strategy for recommending response and recovery actions would be to train a supervised model to map observed security alerts to a set of available actions. A major challenge with this strategy, however, is the need for large amounts of carefully annotated data. If, instead of single actions, the target of interest is a potentially complex *sequence* of actions, the need for rich training data in large quantities is even further accentuated. In practice, unfortunately, the availability of alert data containing detailed labels related to response and recovery actions is often limited. This can be partly attributed to the very nature of the problem itself – relative to the total security data volume, security alerts which lead to actions are rare. More fundamentally, there is still a lack of established and functioning processes for annotating data based on recorded actions taken by human operators [80].

In view of the above challenges, a more reasonable approach to building a response recommendation system is semi-supervised learning, which, in addition to labelled data, makes use of unlabelled data to improve performance in classification and prediction tasks [81, 82]. In recent work, results comparable with state-of-the-art supervised models have been achieved on benchmark problems with semi-supervised approaches, using only a small fraction of the data required by the former (e.g. [83]). For semi-supervised models to be applicable, some basic assumptions about the data are usually made [81]. It is assumed that data points, which form a cluster or lie close to each other in high-density regions, are likely to share the same class label (with an analogous assumption for regression problems). Conversely, the decision boundary between two classes should lie in a low-density region.

For our current task, the implication of the above assumptions is that we must be able to represent security alerts as points in some space, equipped with a notion of distance or similarity between the points. Furthermore, the representation and associated similarity measure must agree with expert judgment, allowing the data to form clusters which are meaningful from a cybersecurity point of view. Concretely, the development of such a representation entails transforming alert data stored as complex JSON objects into a numerical vector with the aforementioned properties.

In addition to being a requirement from the modelling point of view, the ability to measure the similarity between alerts is important and useful and supports several business use-cases. These include supplying information about similar alerts as a tool for security analysts, as well as automatic identification of *false alerts* by comparing with previously documented false positives. While we currently automatically declare alerts to be false positives only based on an exact match, an extension to allow for minor differences between incidents is expected to greatly reduce the number of false positives that require manual reviewing. We note that the automatic declaration of false positives is, in fact, a special case of response recommendation, with the action to be taken being "no action". Trivial, as it may sound, the "no action" decision is a very important one for a security operator to take.

We now proceed with three use cases where similarity evaluation methods were applied to security alerts: preserving the original naming choices, we will speak about (security) *incidents* in the first case and *host aggregations* in the second and third cases. Incidents and host aggregations are produced by two different attack detection mechanisms applied to data points. The data points are collected by security monitoring *sensors* in *hosts* or *endpoints* (these two terms are used in the report interchangeably).

3.1 Incident similarity model

In this section we will describe a method for determining incident similarity in an EDR-style system. In this system, a software agent running locally on each protected endpoint streams events gathered from that system to a back-end system. Events contain information about process launches, filesystem activity, network activity and other low-level operations occurring on the system. Arriving events are processed by a set of rules called detections. Under specific circumstances, an incident may be triggered (the output of a sequence of detections indicated anomalous or potentially malicious behaviour had occurred on that system). An incident is defined as a collection of several related detection events that are generated by detection rules applied to both events collected from endpoints and metadata provided by enrichment processes. Enrichment processes add metadata to information collected from each endpoint, such as the reputation of a URL or IP address, or malware verdicts for known files. From the information contained in a detection, several features are selected as descriptors for the detection. Examples of such features include process name, detection category, detection rule ID, and MITRE ID. The feature values in each detection, all of which are treated as strings, are first collected into a list of tokens and then combined across the entire incident. An incident can contain anywhere between one and many thousands of detections. Incidents containing multiple detections may include multiple of the same detection.

Each processed incident results in a "document" representing that incident. In order to train a machine learning model on this textual data, it must first be converted into a numerical representation. We convert documents into sparse numerical feature vectors using a bag-of-words model. In order to robustly represent incident data despite the differences in numbers of detections, and the presence of identical detections, only unique occurrences of feature values within each incident are considered. Due to the removal of repeating feature values within an incident, the resulting feature vector can be considered binary. A bag of words model simply assigns each unique entry to a vocabulary and then creates a data structure that contains counts of the occurrences of each vocabulary item.

For improved scalability and memory efficiency, we employ a hashing trick [84] to map tokens of the bag-of-words model to elements of {1, 2, …, *N*} for an appropriate choice of *N*. (Note that we omit the "alternate sign" parameter of the hashing method, since we apply **term frequency–inverse document frequency** (tf-idf) to the hashed result and need to ensure that all the resulting vector's components are non-negative.) This hashing method also allows us to avoid storing a dictionary of potentially sensitive data (token names) in memory. To reduce the impact of feature values occurring frequently across all documents, we apply tf-idf weighting to all vector components [85], followed by normalization to the unit Euclidean norm. Term frequency-inverse document frequency is a statistical method often used when processing written languages, where items in the vocabulary are inversely weighted based on how often they appear in a document. Euclidean distances are the straight-line distances measured between two points. The Euclidean norm is the square root of the inner product of a vector with itself.

Finally, the similarity between incidents is computed using the cosine similarity method. Since each incident vector is normalized to the Euclidean norm, computing similarities reduces to computing the inner product between incident vectors. We refer to the trained tf-idf transformation as the *incident similarity model*, owing to its central use-case of evaluating similarities between incidents.

The incident similarity model is retrained periodically on a database of historical data, using a time window of a fixed size. The purpose of the retraining is to keep the model up-to-date as new incidents are recorded, and existing incidents evolve over time (with individual detections being added or removed). This periodical retraining also minimizes concept drift [86]. Here concept drift refers to the fact that observed behaviours (obtained from events on monitored computer systems) change over

time. This can happen when software or the operating system is updated, new software is installed, or because attackers change their tactics to evade existing detection methods.

3.1.1 Model validation and initial results

Upon creation of our initial model, our security experts manually examined similarity scores for many incidents in order to validate that the model was of sufficient quality to be used in production for a limited set of use-cases. As the first use-case, the scoring of incident similarities was made available to security analysts through an API, where incidents could be queried to return a list of the most similar incidents previously encountered within the same organization (see Figure 1). The same information will also be published in the F-Secure Rapid Detection and Response (RDR) service portal in the near future.

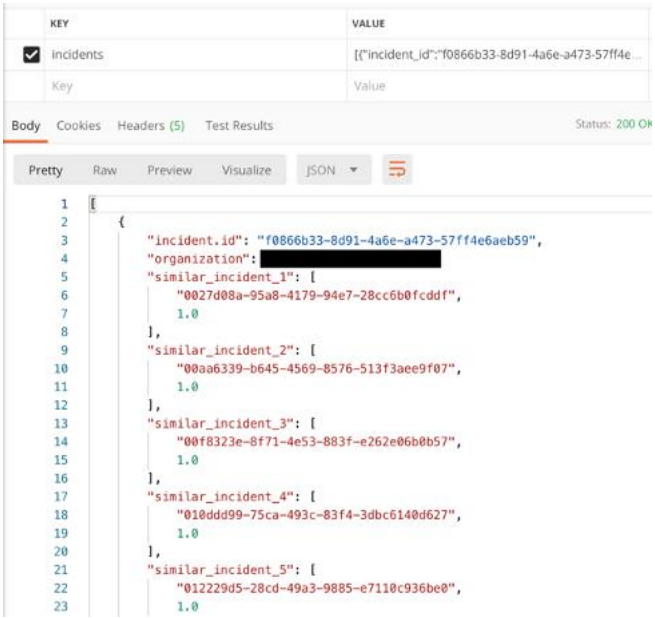

Figure 1. API for making incident similarity queries. The query returns a sorted list of the IDs of the most similar incidents within the same organization, along with the similarity scores. In the displayed example, all the five returned incidents have scores of 1.0, indicating that they are identical to the queried incident.

In order to gauge the utility of the similarity score model, we experimented with comparing the distribution of similarity scores for two different organizations. We trained each model on two months' worth of data and tallied the similarity scores obtained between the training data and one week's worth of query data. The resulting scores were organization-specific, meaning that similarity scores were only calculated between incidents belonging to the same organization. We observed that different organizations had significantly different distributions, with some organizations having distributions concentrated near 0 and 1, while other organizations had a more diverse set of similarity scores between 0 and 1, indicating a more diverse set of security incidents (Figure 2). We hope such experiments can help us characterize the nature of the distribution of security incidents within single organizations and differences between organizations.

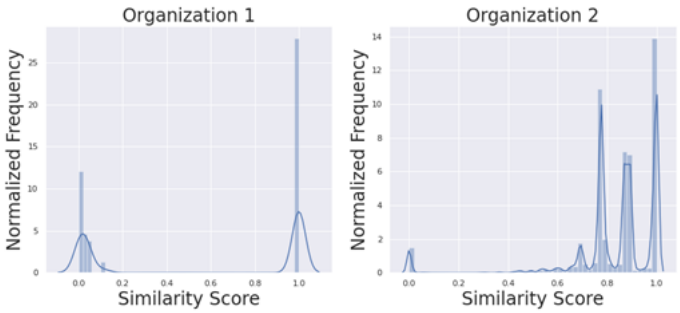

Figure 2: Distribution of incident similarity scores for two separate organizations. One distribution is quite polarized while the other includes many intermediate values, demonstrating the difference in diversity of incidents between the two organizations.

We next investigated whether our incident similarity model might be used as a false positive prevention tool. We posited that if, given a new incident, there are a sufficient number of highly similar incidents in the training data that are already marked as false positive, we may be able to automatically mark the new incident as a false positive. In our current systems, false-positive prevention is implemented based on incident "fingerprints", which demand an exact match between incidents, and is a stronger condition than even having a similarity score of 1.0.

In order to implement false positive prevention, based on the incident similarity model, we first needed to quantify how many new incidents could be declared as already seen by the model. Here the definition of "already seen" depends on two parameters, namely (i) a lower bound n on the number of incidents in the training data with which a new incident matches, and (ii) a lower bound α on the similarity score which is used to declare a match. As we vary these two parameters, we are interested in the fraction of queried (new) incidents that can be declared as having been already seen by the model.

Preliminary results of our experiment with $\alpha = 0.9, 1.0$ and n ranging from 0 to 100 are shown in Figure 3, with a comparison against the currently implemented fingerprint-based false positive prevention as a benchmark. The results are encouraging. For example, for $n = 1$ (we only demand that a new incident has a single match in the training data) we observed that 53% of incidents had at least one match by fingerprint in the training data. However, the similarity score criterion allowed for 72% of new incidents to match with the training data when $\alpha = 1.0$, which increases to 77% when we set $\alpha = 0.9$. As we see more generally, even with the strictest possible α of 1.0, the incident similarity model allows us to increase the fraction of "already seen" new incidents by about 20%. Lowering our standards of what constitutes a match to allow for $\alpha = 0.9$, we gain another 5% - 7% of new incidents that can be declared as already seen by our system.

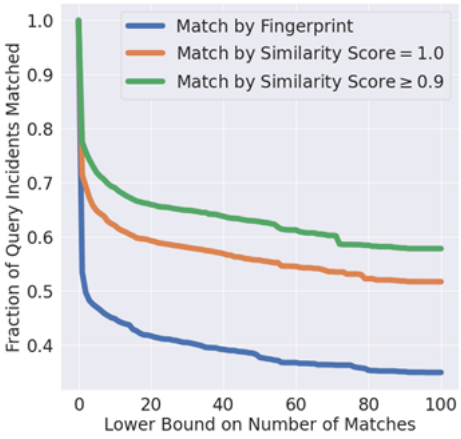

Figure 3: Matching new queried incidents with those in the incident similarity model training data for the purpose of false positive prevention. The incident similarity model allows for the declaration of many more matches than the current algorithm based on identical incident fingerprints.

3.1.2 Conclusions and future work

As a first step towards building an algorithm to recommend response and recovery actions to security operators, we have developed a model based on vector representations of incidents, which allows similarities between incidents to be numerically quantified. The development of this model was done in close collaboration with our security experts. Ongoing improvement of this mechanism will be subject to their iterative manual review and feedback. The fact that the similarity scores produced by our model have been made available both internally to our analysts and externally to our RDR partners (which are typically providers of managed detection and response services) enables us to get feedback from a wide range of users.

While expert feedback is of high importance in developing a sound incident representation, especially in its initial stage of development, we will in future developments additionally put emphasis on more objective evaluation criteria. We will study the clustering properties of the representation to establish its usefulness for semi-supervised learning. Arguably the simplest use-case that can be addressed using a semi-supervised approach is automatic declaration of false positives, based on high similarity to known false positives. Solving this task will pave a way for recommending more complex actions, likely requiring more sophisticated semi-supervised techniques.

3.2 Host aggregation similarity model

While the previously described experiments operated on incidents and detections, the methodology described in this section operates on what we're calling *host aggregations.* Incidents and host aggregations are outcomes of two different event data analysis methods designed to enable different types of attack detection use cases. In order to describe what host aggregations are, we will first describe the data flow and associated cyberattack detection logic which generates them.

Data collection starts at each endpoint, where security sensors collect predefined static and behavioural information. This information includes host configuration data, traces of past and present application activity, and other relevant data. Collected data is transferred into a database server. Inspection operations are carried out regularly to check for signs of attacks in the collected data. These operations are driven by database queries, which serve as atomic detection rules produced by the Managed Detection and Response (MDR) service provider. Each query is given a self-descriptive

name (also referred to as a tag) reflecting what the query looks for. In total, there are several hundreds of tags, and their number varies over time due to day-to-day maintenance changes of the queries (updates, merges, deletions) by security analysts. It is important to emphasize that individual queries usually cannot provide precise and reliable identification of threats and attacks. Instead, a specific query looks for a relatively weak indication of a known suspicious state or activity which may or may not have a connection to a real attack or security breach. Here are a few examples of tag names and their meaning.

- *Ps-arrayobf*: "Powershell command contains potential array obfuscation"
- *Browser-launching-suspicious-proc*: "Browser process launching a suspicious process"
- *Enum-ipconfig*: "Use of the ipconfig command, often used by attackers for information gathering, is detected"

By collecting and consolidating the output of satisfied queries over individual hosts, sets of hosts, organizations, and users, within a defined temporal window, a rich security context is obtained. This context is used for cyber-attack detection logic.

Typically, a single inspection operation applies a few queries (depending on settings such as inspection schedule and last inspection time) that run on endpoint sensor data. The output of an inspection operation is referred to as a *document*, which includes the satisfied queries and the references to the hosts on which those queries triggered. A *document* can thus be considered an organization-wide security scan report.

Documents are then processed in a per-organization fashion, to create different aggregations (e.g. for users and hosts). A host aggregation structure is prepared for each host which includes names of atomic detections (tags), their counts, and references to the documents (scan reports) that assigned these tags. Host aggregations are also utilized for visualization purposes in dashboards. These dashboards are monitored by security analysts in order to discover potentially attacked hosts and to prioritize their routine analysis work.

Documents and host aggregations are represented by JSON structures as follows:

Document format:

{
"id": unique identifier of the document
"endpoints": list of structures representing endpoints included into the document
[...
{ Every structure represents single endpoint with attributes like:
"type": sensor type,
"endpoint_id": endpoint ID
...}
...]
"tags": list of structures representing tags assigned to the endpoints listed above
[...
{Every structure represents one assigned tag with attributes like:
"name": name of tag,
"level": severity level (e.g. HIGH, MEDIUM, LOW and so forth),
"status": status category (e.g. STABLE, TEST and so forth),
"type": type of the tag (e.g. PREPROCESS, NORMAL and so forth),
...}
...]

... other attributes ...
}

Host aggregation format:
{
"id": unique identifier of the host aggregation structure
"endpoint": {...} The structure representing the endpoint (i.e. host) with attributes like: sensor type, endpoint ID (similar to the structure from the "endpoint" list in document layout).
"tags": list of structures representing tags assigned to this endpoint over a time window
[...
{
"name": name of tag
"count": number of assignments from unique documents
"document_ids": {...} The structure representing documents that assigned this tag and their host coverage count
... other attributes for the tag like level, status, type etc.
}
...]
... other attributes ...
}

Manual host aggregation analysis involves combining multiple weak signals together in order to identify whether response operations should be initiated, and if so, what actions should be taken. These manual steps are highly time-consuming and require significant experience. As in our previous section, we aimed to provide functionality to the operators of those systems in order to ease their workload. Our goals were to (i) group similar host aggregation objects over specific organizations together in order to help operators analyze dense groups of such objects by picking a single host aggregation object from the group and (ii) identify dissimilar host aggregation objects (outliers) that could be treated as high-priority anomalous cases. The research goal can thus be considered a clustering and outlier detection problem. A relevant example can be found in [87].

Our experiments were carried out on a sample set containing hundreds of thousands of host aggregations, produced over a time window of 24 hours, for several hundred organizations. The numbers of unique documents referred to by the host aggregations for specific organizations varied between several thousands and several hundreds of thousands. The number of unique tags per organization was in the order of hundreds.

As in our previous experiment, we first needed to find a way to convert string-based data representations into numeric values. Looking at the data structure above, we can see that both documents and host aggregations can contain lists of hosts. Host aggregations can also contain lists of documents. These somewhat complex data structures needed to be converted into simpler representations based on individual hosts. If we define each host's feature vector by the presence of tags within the scope of a specific organization, it should be straightforward to build relatively low-dimensional sparse vectors of integers (tag counts) for each host. However, each document or host aggregation structure can contain many other fields and creating vector representations containing all those fields would generate very high-dimensional feature vectors. After checking with the security analysts who work directly with this data, we were able to identify several irrelevant fields that could be omitted, and thus significantly reduce the final dimensionality of our vectors. Our final design used one row per host aggregation entry, with boolean features (columns) to indicate the presence of tag-documentID pairs. Term frequency–inverse document frequency (tf-idf) transformation was ap-

plied to the matrix elements in order to reduce the influence of the feature values which were encountered in large numbers of host aggregation objects. The resulting matrix for the largest organization was several hundreds of thousands of rows by several hundreds of thousands of columns.

Sparse high dimensional data is often clustered with dimensionality reduction methods such as PCA, t-SNE, or UMAP. However, in order to ensure better interpretability of intermediate decisions, we opted, at least for the first phase of the research, to avoid such methods, and use density-based spatial clustering of applications with noise (DBSCAN) introduced by Ester et al. in [88]. We selected two different distance-measurement methods, based on the size of the organization. For smaller and medium-sized organizations, we used Jaccard distance, and for large organizations we used Euclidean distance. For DBSCAN parameters, we set the minimal cluster size to 2, which both allowed for identification of outliers, and made it possible to "rank" host aggregation objects by the size of the clusters they belong to. The DBSCAN hyperparameter defining the maximum acceptable distance between two similar (belonging to the same cluster) objects is an adjustable parameter that needs to be tuned after receiving feedback from security analysts.

3.2.1 Model validation and initial results

We integrated our mechanism into a staging pipeline responsible for data visualization dashboards via a shared Python library. While we ultimately plan on establishing a continuous feedback loop with security operators to assess the value of our clustering method, for the time being we are validating it by examining the "compression rate" of the host aggregation objects set (Figure 4).

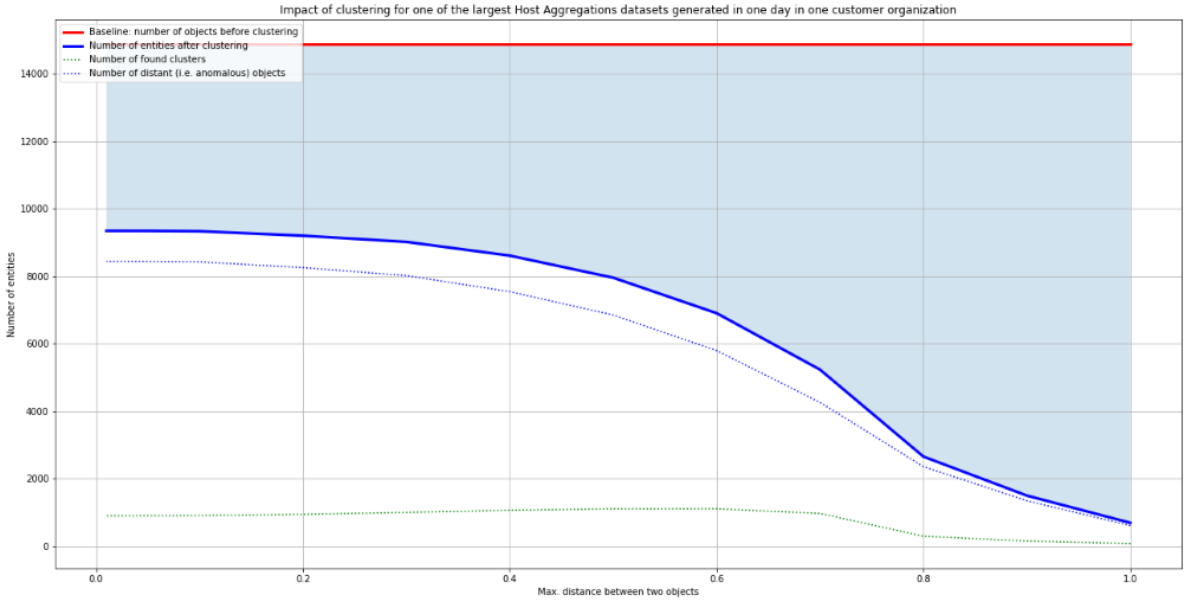

Figure 4. Example of applying the current approach (DBSCAN-specific)

Figure 4 presents the number of unique entities (that is, clusters and outliers – distant "anomalous" host aggregation objects) obtained for different values of the "maximum acceptable distance between two similar points" parameter in DBSCAN. The solid red line depicts the original number of host aggregation objects, and the solid blue line shows the final number of the unique entities after clustering (the sum of the number of clusters and the number of anomalous distant host aggregations) for a given maximum acceptable distance value. The dotted blue and green lines depict the numbers of discovered outliers and clusters respectively. The intuition here is that via clustering we want to reduce the number of entities that security analysts must inspect, assuming only one host aggregation object from a cluster must be inspected. This assumption, of course, requires sufficiently

tight clusters, so there is an obvious trade-off between the total number of clusters and outliers and the tightness of the clusters, which is controlled by the maximum acceptable distance parameter of DBSCAN and can be tuned to reflect the security operators' preferences.

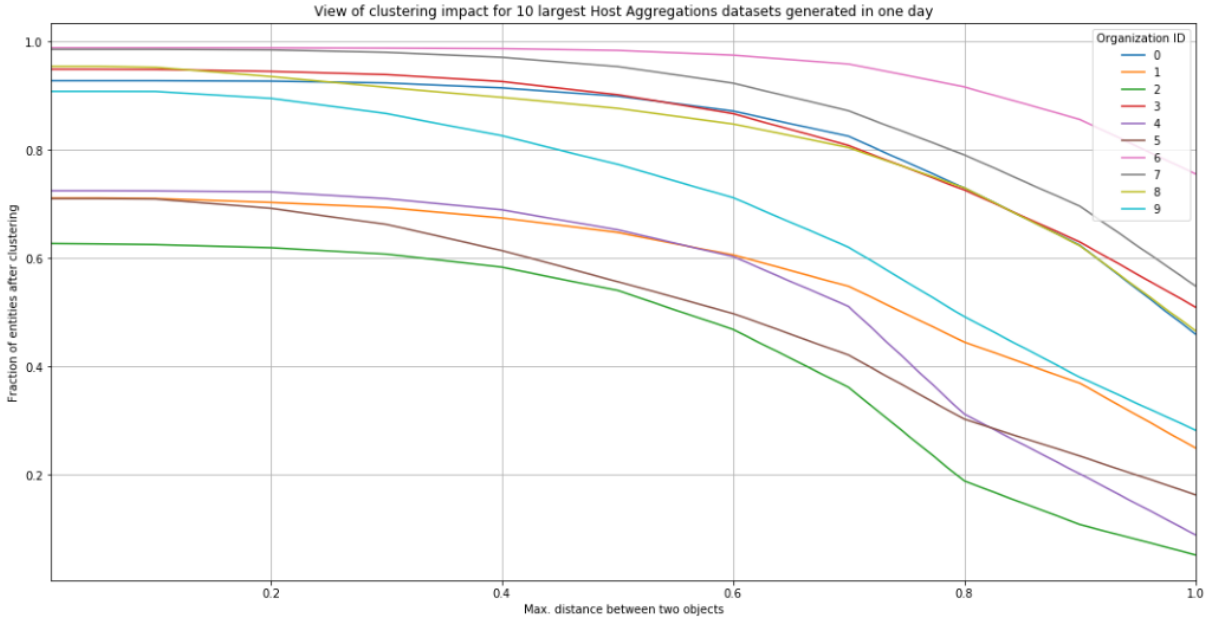

Figure 5. Comparison of the impact of clustering for different organizations

In Figure 5, we can see that the impact of clustering can vary significantly across organizations. In particular, the clustering brings little value in the case of organization 6 and is dramatically more effective for organization 2. Our investigation shows that this is likely a consequence of the differences in the frequency of inspection operations carried out in specific organizations and, thus, in the number of distinct documents generated by those operations per day. This observation naturally leads to a question of optimal strategies for running inspection operations, clearly an interesting and challenging one.

3.2.2 Conclusions and future work

Our current approach to dealing with host aggregation data generated by our Managed Detection and Response (MDR) service provider relies on using unlabeled data to construct dense groups and highlight anomalous host aggregations for each organization. We see two key directions to develop the approach further after establishing a reliable process to receive feedback from security analysts:

- For the clustering track, we have open questions about fine-tuning of the suggested approach (in particular, how the maximum acceptable distance and minimal cluster size values are to be defined) that could lead us to next steps of considering alternative clustering methods such as (i) using the host aggregations' data across multiple or all organizations; (ii) supporting a streaming processing mode; and (iii) using dimensionality reduction approaches to make compact, dense representations of the data before passing it to clustering methods.
- When it comes to the anomaly detection track, the expected operator feedback will contain outliers highlighting host aggregations that are relevant and irrelevant for cyberattack detection. This information will help us revisit our current approach in the semi-supervised or supervised setting [89].

3.3 False Alert recognition

In this section, we will detail a slightly different false alert prevention mechanism built on top of the same host aggregation system described in the previous section. As already mentioned, security analysts monitor host aggregation data using various dashboards, in search of potential attacker behaviour. If an incident is spotted, the analyst creates a ticket in a JIRA tracking system. JIRA tickets include information about the host and its context, including host aggregation tags. Other security analysts follow these tickets and process them based on importance. When processing an issue, an analyst will examine available data in order to determine whether the incident was real, or a false alert. Such an investigation may also include collection of data from elsewhere. After the investigation the analyst records their verdict in the JIRA ticket. This section details methodology we developed to train classifiers on these JIRA tickets, in order to automatically recognize false alerts.

When considering the features of alerts (and perhaps other contextual information) and examples of security analyst decisions on the relevance of those, as contained in these tickets, the problem turns into one of classification, for which several machine learning techniques have been developed. For such a classifier to be useful, it is not necessary to reach a very high classification accuracy - it is enough to compute a score reflecting the probability of an alert being a true or false positive and provide the score together with the alert. Subsequently, alerts can be presented in descending order, according to their scores, in the security analyst's user interface. This provides the analysts with a prioritized list of alerts to investigate.

In order to build a mechanism capable of classifying incidents in JIRA tickets, we first needed to harvest relevant information from those tickets in order to create a dataset with which to build our model. In this case, the process was rather easy – we selected JIRA tickets that included host aggregation tags and used those tags as features to represent each issue. Host aggregation tags were present in most of the tickets, their semantic information was the same across all issues, and they were easy to process out. We used verdicts in the tickets, provided by analysts, as ground truth and the target for the prediction task. The sample set contained around 10,000 tickets, of which about 10% were true positives (i.e. real threats) and the rest false alerts (something out of the ordinary happening on a system that turned out not to be malicious).

Data pre-processing was critical to the quality of the results of this experiment. We needed to ensure that the information in the target variable (analyst verdict in the issue) was not somehow accidentally encoded in the variables used for prediction. As the training data had quite a simple structure, and we included only counts of tags, pre-processing was straightforward. We chose to scale all variables to zero mean and unit variance in order to reduce the effect of features with very high values. Scaling is also necessary for the L2 regularization method we used during modelling. Although this is not necessarily needed for tree-based classification methods, it is still a good practice.

Tickets contained a variety of verdicts, including False Positive, Red Teaming, Malware, and Compromised Account. We chose to limit labels in our dataset to just two by mapping all but "False Positive" to the same label. Fine-grained information about the type of true positive was not required. Converting the classification task to a binary case made model fitting easier and likely made our predictions more reliable.

3.3.1 Model validation and initial results

Due to the simplicity of our dataset, we tried a few classification models, including Logistic Regression, Random Forest and Gradient Boosting, and compared results using ROC curve and AUC values. See Table 1 for a comparison of model performances and Figure 6 for an illustration of the ROC curves between the compared models. The differences in performance were modest, and we ended

up choosing the Logistic Regression model due to its simplicity. Only linear terms were considered, and all higher powers of the features and cross terms were discarded.

Algorithm	AUC value	F1 score	Accuracy
Logistic Regression (LR)	0.71	0.23	0.93
Random Forest (RF)	**0.84**	**0.39**	**0.95**
Gradient Boosting (XG)	0.75	0.34	0.87

Table 1. Summary of performances for different models.

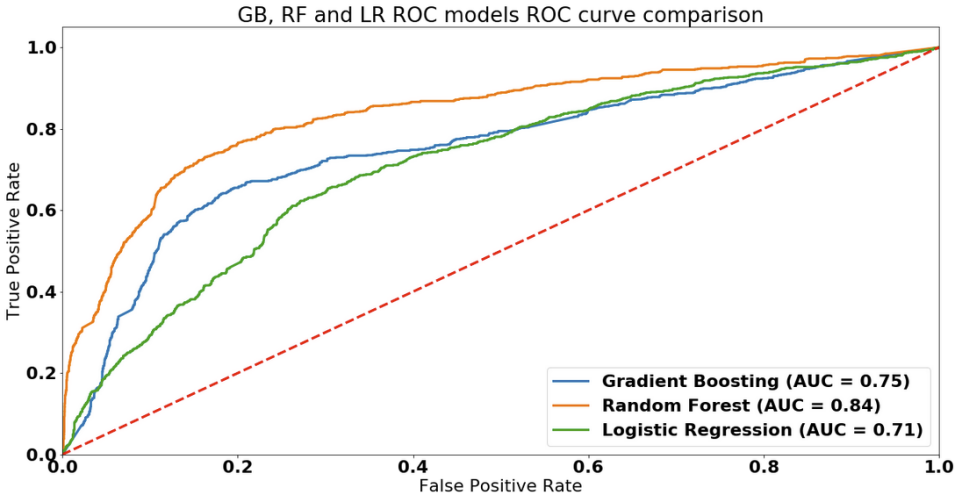

Figure 6. ROC visualization for all candidate models.

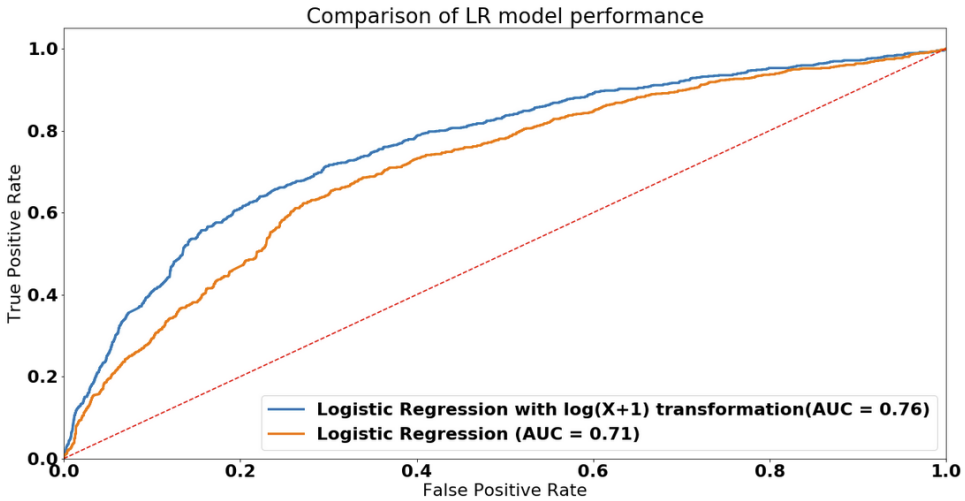

Figure 7. ROC visualization with transformed and non-transformed data.

It should also be noted that the F1 scores show only limited success in the classification task, whereas the accuracies are very high. This is due to the highly imbalanced true and false positive class distribution, and partially due to noise in the data – similar host aggregations can be interpreted differently with relation to the host (i.e. the same set of tags can be considered suspicious for some hosts, whereas for other hosts such behaviour can be normal). As stated before, perfect classification is not required to make this approach applicable – a reliable probability of the case being a true positive is sufficient to give the alerts a priority order. The reliability of the probability provided by our model can be shown both from the ROC curve, and from Figure 8 which depicts how the predicted probabilities are distributed amongst the ground truth classes. From Figure 8 one can observe that most of incidents with a predicted probability of over 0.20 are likely actual true positive cases, whereas values below this are likely false positive cases. There is some overlap when the predicted

value is between 0 and 0.2, and our current work focuses on finding distinguishing features for these cases.

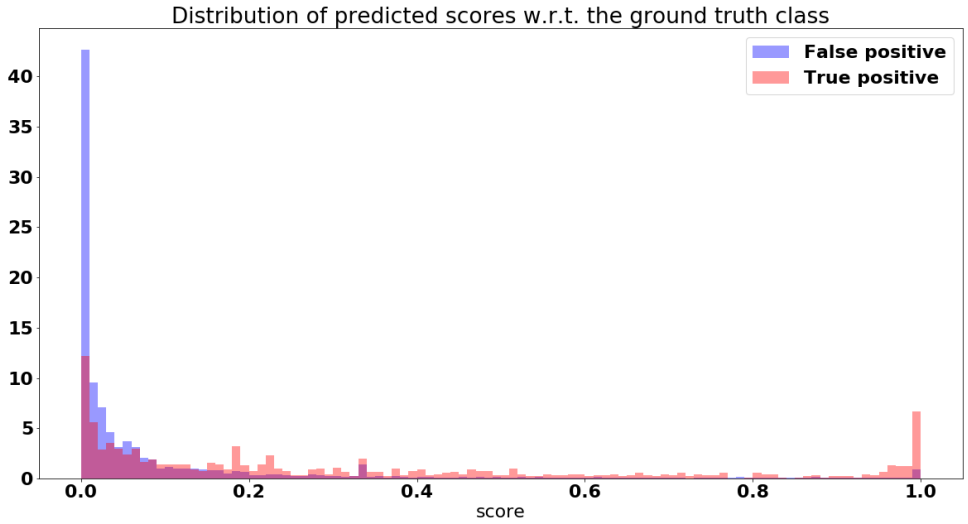

Figure 8. Distribution of predicted scores.

All model hyperparameters were determined using grid search and five-fold cross validation. For the chosen logistic regression model, the only hyperparameter we needed to pay extended attention was the L2 regularization parameter, as this parameter can have drastic effect on the model – too high regularization will cause underfitting, whereas too low regularization will cause the model to overfit. We used negative logarithmic loss as the criteria for the parameter tuning. When optimal parameters were found, this model was fitted to all the available data.

Implementation of a method to prioritise tickets to be handled by human specialists includes a pathological feedback loop. If the prioritisation is strict and only issues predicted to be true positives with high probability are ever checked, cases that have lower probability will never be checked. This in turn leads to bias in the data that will be used to fit the next prediction model. As low probability true positives do not get checked by specialists, the training data will not contain labelled examples of such, and the feedback mechanism will increasingly drive the model to classify samples that are similar to the true positives present in the initial data set, thus preventing the discovery of new incident types. Mechanisms need to be implemented in the production system to break this feedback loop. The simplest solution is to always take a random sample from issues that are presented to analysts, regardless of the model prediction. Ideally the new model would then be trained on the verdicts given to samples in this random set to prevent bias. In practice this may be suboptimal, and a compromise where both random and high probability samples are used in the next training set may be the preferred solution. This question remains open.

3.3.2 Conclusions and future work

Our initial work for validating this approach can be considered complete. Even though we are using a simple model to classifying issues, the methodology is sufficiently accurate, indicating that this task can be reliably modelled. The next phase is to put the model into operational use in our intrusion detection system in order to provide true practical value for analysts and assess its potential. This task includes implementing scoring logic, integrating the model into our detection flow, and implementing a method to easily fit the model to new data when it becomes available. The process of triggering a model update will be manual at first and then scheduled once everything is in working order.

While putting our model into production, we intend to improve it. Improvement ideas currently center around the use of other features in addition to the host aggregation tags, such as the context of the issue (host, organisation, user), time series aspects of the host (what happened previously on a given host), and global context (known ongoing incidents on other hosts in the same organisation).

4 Conclusions

Both security vendors and the European Union are putting a great deal of effort and resources into researching mechanisms to intelligently automate tasks in breach detection and incident response workflows. One of the most important of those tasks is to find accurate methods to reduce false positives and discover incident similarity. The less noise a security analyst needs to deal with, the more likely it is that they'll find real incidents. As part of this effort, SAPPAN is actively conducting research and developing machine learning-based approaches to solve these problems.

Results from our research demonstrate that, even when using simple techniques and a limited amount of labelled data, it is possible to train models that provide tangible benefits to security analysts. Although further iteration and improvement of these techniques is still required, we are pleasantly motivated with our initial results.

5 Acknowledgements

We would like to thank Gabriela Aumayr, Josef Niedermier and Scott Dolin from HPE for sharing their view of such a complex industry. We would also like to thank Mischa Obrecht, Cybersecurity Specialist, Dreamlab Technologies AG, for contributing his insight into realities of providing cybersecurity services. Finally, we would like to thank all the SAPPAN consortium members, Matti Aksela and the Artificial Intelligence Center of Excellence team at F-Secure for support, comments and helpful discussions.

6 Bibliography

1. https://enterprise.verizon.com/resources/reports/2020-data-breach-investigations-report.pdf
2. https://en.wikipedia.org/wiki/Endpoint_Detection_and_Response
3. https://blogs.gartner.com/anton-chuvakin/2013/07/26/named-endpoint-threat-detection-response/
4. https://www.gartner.com/doc/reprints?id=1-1OA93LZ7&ct=190716&st=sb
5. https://www.gartner.com/en/information-technology/glossary/security-information-and-event-management-siem
6. https://docs.broadcom.com/doc/endpoint-security-en
7. https://en.wikipedia.org/wiki/Framework_Programmes_for_Research_and_Technological_Development#Horizon_2020
8. https://www.mcafee.com/enterprise/en-us/products/mvision-edr.html
9. https://players.brightcove.net/21712694001/S1o50VS1l_default/index.html?videoId=6065865044001
10. https://www.mcafee.com/enterprise/en-us/assets/data-sheets/ds-mvision-edr.pdf
11. https://www.trendmicro.com/en_us/business/products/detection-response/edr-endpoint-sensor.html

12. https://www.trendmicro.com/en_us/business/products/detection-response/edr-endpoint-sensor.html?modal=s3b-icon-datasheet-762c57
13. https://www.trendmicro.com/en_us/business/products/detection-response/edr-endpoint-sensor.html?modal=s3a-icon-esg-tm-video-873ecb
14. https://www.sophos.com/en-us/products/endpoint-antivirus/edr.aspx
15. https://www.sophos.com/en-us/products/endpoint-antivirus/tech-specs.aspx
16. https://news.sophos.com/en-us/2016/10/28/watch-now-sophos-intercept-x-root-cause-analysis-in-two-minutes/
17. https://docs.sophos.com/central/MTR/welcomeguide/sophos-managed-threat-response-welcome-guide.pdf
18. https://www.sophos.com/en-us/press-office/press-releases/2019/01/sophos-acquires-darkbytes-as-foundation-of-new-mdr-services.aspx
19. https://www.atarlabs.io/en/
20. https://www.atarlabs.io/en/static/automate-repetitive-activities/
21. https://ayehu.com/
22. https://www.eclecticiq.com/
23. https://www.eclecticiq.com/platform/features
24. https://www.splunk.com/en_us/software/splunk-security-orchestration-and-automation.html
25. https://splunkproducttours.herokuapp.com/tour/splunk-phantom
26. https://www.splunk.com/en_us/software/splunk-security-orchestration-and-automation/features.html
27. https://www.pandasecurity.com/rfiles/enterprise/documentation/fusion360/FUSION360-datasheet-EN.PDF
28. https://www.pandasecurity.com/rfiles/newhome2016/microsite AD/resources/Datasheets/Adaptive_Defense-Advanced_Reporting_Tool-en.pdf
29. https://www.carbonblack.com/products/enterprise-edr/
30. https://www.carbonblack.com/products/endpoint-standard/
31. https://www.carbonblack.com/products/edr/
32. https://www.crowdstrike.com/endpoint-security-products/falcon-x-threat-intelligence/
33. https://www.crowdstrike.com/endpoint-security-products/falcon-insight-endpoint-detection-response/
34. https://www.youtube.com/watch?v=j9761pD0X3A
35. https://www.youtube.com/watch?v=cilLVl7IKwc
36. https://logrhythm.com/
37. https://logrhythm.com/press-releases/logrhythm-granted-patent-for-technological-strides/
38. http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=10,091,217.PN.&OS=PN/10,091,217&RS=PN/10,091,217
39. https://www.youtube.com/watch?v=5x-KRnCw-Y4
40. https://www.youtube.com/watch?v=jNqr5Cvh3pA&list=PLxBs-yP4NjaOuh4aCrrKPFi9cG7awAqjG
41. https://s7d2.scene7.com/is/content/cylance/prod/cylance-web/en-us/resources/knowledge-center/resource-library/briefs/AIDrivenThreatPreventionandEDRSolutionBrief.pdf
42. https://www.netscout.com/product/arbor-sightline-sentinel
43. https://www.netscout.com/product/arbor-sightline-sentinel#click-watch-video
44. https://www.netscout.com/sites/default/files/2019-09/SECPDS_004_EN-1901%20-%20Arbor%20Threat%20Mitigation%20System%20%28TMS%29.pdf
45. https://respond-software.com/
46. https://d53g0hkpcf8eh.cloudfront.net/wp-content/uploads/Respond-Software-Respond-Analyst-Datasheet.pdf

47. https://d53g0hkpcf8eh.cloudfront.net/wp-content/uploads/2019/01/Respond-SoftwarePalo-Alto-Networks-JointSolutionBrief.pdf
48. https://www.servicenow.com/products/security-operations.html
49. https://www.servicenow.com/now-platform.html
50. https://www.servicenow.com/products/predictive-intelligence.html
51. https://www.ibm.com/products/cognitive-security-analytics
52. https://www.ibm.com/downloads/cas/52GBXLK8
53. https://www.ibm.com/security/intelligent-orchestration/resilient
54. https://www.elastic.co/security
55. https://www.darktrace.com/en/products/enterprise/
56. https://www.darktrace.com/en/resources/wp-antigena.pdf
57. https://www.paloaltonetworks.com/cortex/detection-and-response
58. https://start.paloaltonetworks.com/deploying-cortex-in-soc.html
59. https://www.paloaltonetworks.com/cortex/xsoar
60. https://start.paloaltonetworks.com/deploying-cortex-in-our-own-soc-on-demand-success.html
61. https://go.demisto.com/hubfs/Resources/Datasheets/data_sheet_final.pdf
62. P. Nespoli, D. Papamartzivanos, F. Gómez Mármol and G. Kambourakis, "Optimal Countermeasures Selection Against Cyber Attacks: A Comprehensive Survey on Reaction Frameworks," in IEEE Communications Surveys & Tutorials, vol. 20, no. 2, pp. 1361-1396, Secondquarter 2018, doi: 10.1109/COMST.2017.2781126.
    a. https://ieeexplore.ieee.org/document/8169023
63. Nyre-Yu M., Sprehn K.A., Caldwell B.S. (2019) Informing Hybrid System Design in Cybersecurity Incident Response. In: Moallem A. (eds) HCI for Cybersecurity, Privacy and Trust. HCII 2019. Lecture Notes in Computer Science, vol 11594. Springer, Cham
    a. https://doi.org/10.1007/978-3-030-22351-9_22
64. Andrade R., Torres J., Cadena S. (2019) Cognitive Security for Incident Management Process. In: Rocha Á., Ferrás C., Paredes M. (eds) Information Technology and Systems. ICITS 2019. Advances in Intelligent Systems and Computing, vol 918. Springer, Cham
    a. https://link.springer.com/chapter/10.1007%2F978-3-030-11890-7_59
65. https://arxiv.org/pdf/1903.10080v1.pdf
66. M. Yun, Y. Lan and T. Han, "Automate incident management by decision-making model," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, 2017, pp. 217-222, doi: 10.1109/ICBDA.2017.8078811.
    a. https://ieeexplore.ieee.org/document/8078811
67. N. Gupta, I. Traore and P. M. F. de Quinan, "Automated Event Prioritization for Security Operation Center using Deep Learning," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 5864-5872, doi: 10.1109/BigData47090.2019.9006073.
    a. https://ieeexplore.ieee.org/document/9006073
68. M. Husák and J. Kašpar, "Towards Predicting Cyber Attacks Using Information Exchange and Data Mining," 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), Limassol, 2018, pp. 536-541, doi: 10.1109/IWCMC.2018.8450512.
    a. https://ieeexplore.ieee.org/document/8450512
69. HUSÁK, Martin a Jaroslav KAŠPAR. AIDA Framework: Real-Time Correlation and Prediction of Intrusion Detection Alerts. In Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019). New York: ACM, 2019. s. "81:1-81:8", 8 s. ISBN 978-1-4503-7164-3.
    a. https://is.muni.cz/repo/1540256/2019-CyberTIM-AIDA-Framework-paper.pdf
70. Bartos, Vaclav & Zadnik, Martin & Habib, Sheikh & Vasilomanolakis, Emmanouil. (2019). Network entity characterization and attack prediction.
    a. https://arxiv.org/pdf/1909.07694.pdf

71. Bartoš, Václav. (2019). NERD: Network Entity Reputation Database. ARES '19: Proceedings of the 14th International Conference on Availability, Reliability and Security. 1-7. 10.1145/3339252.3340512.
    a. https://nerd.cesnet.cz/
72. SOCCRATES: https://cordis.europa.eu/project/id/833481
73. SOCCRATES: https://www.soccrates.eu/
74. CyberSANE: https://cordis.europa.eu/project/id/833683
75. CyberSANE: https://www.cybersane-project.eu/
76. SPARTA: https://cordis.europa.eu/project/id/830892
77. SPARTA: https://www.indracompany.com/en/indra/sparta-strategic-programs-advanced-research-technology-europe
78. REACT: https://cordis.europa.eu/project/id/786669
79. REACT: https://react-h2020.eu/
80. Shahid Anwar, Jasni Mohamad Zain, Mohamad Fadli Zolkipli, Zakira Inayat, Suleman Khan, Bokolo Anthony, and Victor Chang. From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions. Algorithms, 10(2):39, 2017.
81. Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-supervised learning*. MIT Press, London, UK, 2006.
82. Feng Liang, Sayan Mukherjee, and Mike West. The use of unlabeled data in predictive modeling. *Statistical Science*, 22(2):189–205, 2007.
83. Yunus Saatchi and Andrew G Wilson. Bayesian GAN. In *Advances in Neural Information Processing Systems*, pages 3622–3631, 2017.
84. Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120, 2009.
85. Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.
86. João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys* (CSUR), 46(4):1–37, 2014.
87. Leung, K. and Leckie, C., 2005, January. Unsupervised anomaly detection in network intrusion detection using clusters. In Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38 (pp. 333-342).
88. Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996, August. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).
89. Görnitz, N., Kloft, M., Rieck, K. and Brefeld, U., 2013. Toward supervised anomaly detection. Journal of Artificial Intelligence Research, 46, pp.235-262.