An Enhanced Cyber Attack Attribution Framework

Nikolaos Pitropakis[1], Emmanouil Panaousis[2],
Alkiviadis Giannakoulias[3], George Kalpakis[4],
Rodrigo Diaz Rodriguez[5], and Panayiotis Sarigiannidis[6]

[1] Edinburgh Napier University
[2] Surrey Centre for Cyber Security, University of Surrey
[3] European Dynamics SA
[4] Information Technologies Institute, Centre for Research and Technology Hellas
[5] Atos Spain SA
[6] University of Western Macedonia

Abstract. Advanced Persistent Threats (APTs) are considered as the threats that are the most challenging to detect and defend against. As APTs use sophisticated attack methods, cyber situational awareness and especially cyber attack attribution are necessary for the preservation of security of cyber infrastructures. Recent challenges faced by organizations in the light of APT proliferation are related to the: collection of APT knowledge; monitoring of APT activities; detection and classification of APTs; and correlation of all these to result in the attribution of the malicious parties that orchestrated an attack. We propose the E**n**hanced Cyb**e**r Attack Attributi**on** (NEON) Framework, which performs attribution of malicious parties behind APT campaigns. NEON is designed to increase societal resiliency to APTs. NEON combines the following functionalities: (i) data collection from APT campaigns; (ii) collection of publicly available data from social media; (iii) honeypots and virtual personas; (iv) network and system behavioural monitoring; (v) incident detection and classification; (vi) network forensics; (vii) dynamic response based on game theory; and (viii) adversarial machine learning; all designed with privacy considerations in mind.

1 Introduction

The financial crisis made Information Technology (IT) infrastructures around the world divert their business plans and often reduce expenditures. Although these reductions have not been reflected on the productivity line, they did however, affect cybersecurity. At the same time, malicious parties have advanced their technology and have managed to be one step ahead of those who try to defend their infrastructures. In 2010, Stuxnet's identification totally reshaped the cybersecurity landscape along with the perception about cyber threats. Advanced Persistent Threats (APTs) had already made a statement, and they stood up to their name with Duqu in 2011, Flame in 2012, Red October in 2012 and MiniDuke in 2013. All of those attacks impacted critical infrastructures.

During the past decade, large-scale and well-organized cyber attacks have become more frequent. In 2017, the Shadow Brokers hacking group came up with a Windows platform exploit named as EternalBlue. This was later used as a part of the WannaCry ransomware that affected numerous countries around the world and their critical infrastructures such as the UK's National Health System (NHS), where it proved to have devastating and life threatening effects, resulting in delays of treatments of patients who were suffering from long term illnesses.

Motivation. Although, the cybersecurity and scientific communities have developed several defensive mechanisms against APTs, there is a number of different challenges that have not been fully addressed. First of all, there is scattered information about APT campaigns, hidden in technical reports as well as in scientific publications that has neither been collected nor visualized in order to facilitate a potential exchange of intelligence. In detail, the sources contain lots of valuable information e.g., domain names, IPs and malware hexes, which have been used in each APT campaign. The same sources contain useful elements that can lead to the detection of a lot of social engineering attacks of which their main target is the human factor. The latter has not been taken into consideration yet when it comes to augment the capabilities of honeypots [1]. In addition to that, malicious parties often reveal information about their activities through social media, which can contribute another valuable source of information.

Conventional incident detection and classification mechanisms have to face a new threat that of adversaries who aims to harm defending mechanisms that use machine learning introducing a new field of research called adversarial machine learning [2]. As malicious parties become aware of the machine learning techniques used in defensive strategies they become elusive, lowering the accuracy rate of all detection capabilities. All of those identified issues are immediately connected with two pillars on which cybersecurity community should depend on; *attribution* and *cybersecurity situational awareness*. The first reflects the need to identify who (i.e., cyber attacker) is responsible for the orchestration of a cyber attack. Like police processes use every piece of evidence coming from investigation and forensic science to understand who are responsible for an incident and their motivating factors, cybersecurity science has exactly the same need. The identification of the malicious parties who have orchestrated large-scale cyber attacks and their correlation with former activities can greatly impact the timing and efficiency of their detection. Additionally, as social engineering attacks take advantage of the human factor, which is referred as the weakest link [3], cybersecurity situational awareness must increase towards protecting cyber infrastructures.

Our contribution. We introduce the E**n**hanced Cyb**e**r Attack Attributi**on** (NEON) Framework, illustrated in Fig. 1, which is designed to accommodate components that address the aforesaid challenges. NEON leads to a user-centric automated cybersecurity platform that gathers heterogeneous data coming from APT reports and publicly available information from social media. By using this material as ground truth, NEON correlates this with other data collected from network and system behavioural monitoring components. To increase defence against social engineering attacks, NEON uses honeypots that attract the attention of potential attackers through the creation and management of virtual personas [1]. The virtual personas accelerate the manifestation of the attacks in contained environments, drawing at the same time valuable information about the adversaries. As part of dynamic response against APTs, NEON uses a game theoretic approach to propose optimal cybersecurity actions against them. All the above result in an integrated system of early detection, classification, optimal response and attribution of APTs. To the best of our knowledge, *NEON is the first framework that has been designed with the ultimate goal to perform enhanced attribution of APT campaigns*. We envisage that the implementation of NEON will have great impact to the situational awareness of cyber infrastructures against sophisticated cyber attacks.

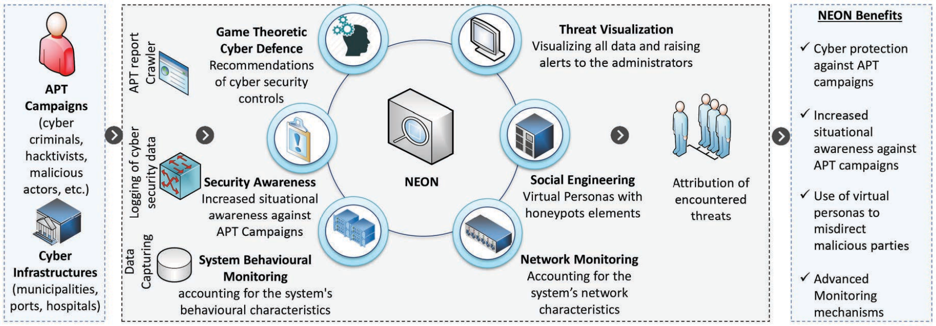
APT Campaigns (cyber criminals, hacktivists, malicious actors, etc.)
Cyber Infrastructures (municipalities, ports, hospitals)
APT report Crawler
Logging of cyber security data
Data Capturing
Game Theoretic Cyber Defence
Recommendations of cyber security controls
Threat Visualization
Visualizing all data and raising alerts to the administrators
Security Awareness
Increased situational awareness against APT Campaigns
NEON
Social Engineering
Virtual Personas with honeypots elements
System Behavioural Monitoring
accounting for the system's behavioural characteristics
Network Monitoring
Accounting for the system's network characteristics
Attribution of encountered threats
NEON Benefits
✓ Cyber protection against APT campaigns
✓ Increased situational awareness against APT campaigns
✓ Use of virtual personas to misdirect malicious parties
✓ Advanced Monitoring mechanisms

Fig. 1: NEON Approach.

Outline. The rest of the paper is organized as follows: Section 2 offers some background information about APTs; Section 3 provides a related literature review; Section 4 introduces the NEON framework and briefly describes its components; Section 5 describes NEON's operation in a healthcare usecase; and finally, Section 5 draws the conclusions giving some pointers for future work.

2 Background

After 2010 and Stuxnet's identification a new terminology was introduced by cybersecurity experts, that of Advanced Persistent Threat (APT) [4]. Advanced, because the adversary is conversant with computer intrusion tools and techniques and is capable of developing custom exploits; Persistent, because the adversary intends to accomplish a mission. They receive directives and work towards specific goals; Threats, because the adversary is organized, funded and motivated. An APT is a multi-step attack designed to infiltrate a system and remain there undetected for a long period of time to obtain high-value information. A characteristic of APTs is that they may spend a significant interval of time between different attack stages. In addition, an APT may combine different attacks types, e.g., zero-day attacks (exploitation of unpatched vulnerabilities) and advanced social engineering attacks. In 2009, when Stuxnet was created, multiple APT campaigns have been identified, e.g., Duqu in 2011, Flame in 2012, Red October in 2012 and MiniDuke in 2013. From 2013 on-wards, the frequency of identified APTs has greatly increased. This is reflected on the posts of major security software companies that have published numerous reports regarding these threats [5], [6], and [7].

APTs' number one target has always been organizations with high value assets hence the reason behind the persistency of those attacks. The information obtained from APTs can be used for an active (i.e., with immediate disruption) or passive (i.e., reconnaissance) malicious action. APTs usually cause major data breaches or they are part of cyber-espionage to cripple critical cyber-physical infrastructures. While APTs can impact political agendas, military plans, and government operations as well as enterprise operations and revenues, the most concerning scenarios are attacks that impact critical infrastructures, such as an electric power grid or water or fuel operations [8].

According to [9], "the attribution problem, which refers to the difficulty of identifying those initially responsible for a cyber attack and their motivating factors, is a key in solidifying the threat representation", while [10] states that "attribution of cyber attacks is not a straight-forward task". Attribution has become an area of research interest the past few years as the attacks towards cyber infrastructures have increased in terms of frequency and impact. So far, there is no concrete methodology that attributes each attack to the malicious parties who launched it. Additionally, no methodology takes into consideration past knowledge of APT campaigns and both network and system behavioural data.

3 Related Work

Advanced Persistent Threats. Several attempts to track, disable or counter APTs have been proposed. Giura et al. [11] propose a Context-Based Detection Framework that introduces the attack pyramid model and takes into account all events occurred in an organization. Specifically, their methodology correlates all relevant events across all pyramid planes, to detect an APT within a specific context as initially collected events are classified into contexts. Virvilis and Gritzalis [12] dive deeper into the technical reports of Stuxnet, Flame and Red October proposing potential countermeasures and defences against APT campaigns. They discuss patch management, strong network access control and monitoring, the importance of Domain Name System related to Command and Control (C&C) servers, protocol-aware security, and usability of Host Based Intrusion Detection Systems along with honeypots. Roman et al. [13] elaborate on the honeypot use and propose solutions that lead to the detection of APTs as honeypots can outperform ordinary solutions contributing to the identification of zero-day exploits.

By performing a study in APTs, Chen et al. [14] refer to potential countermeasures that besides traditional defense mechanisms, advanced malware detection, event anomaly detection and data loss prevention, they point out the need for security awareness and training and intelligence-driven defense. They also discuss the usability of the proposed countermeasures. In [15], Friedberg et al. propose an anomaly detection technique for APTs based on both network and system behaviour while Marchetti et al. [16] narrowed down the problem of detecting APT activities through network monitoring by making use of high volumes of network traffic and ranking the most suspicious internal hosts, which allows security specialists to focus their analyses on a small set of hosts out of the thousands of machines that typically characterize large organizations.

From another point of view [17], Hu et al. consider the joint threats from APT attacker and the insiders, and characterize the interplay as a two-layer game model, i.e., a defense/attack game between defender and APT attacker and an information-trading game among insiders. Consequently, they use game theoretic models to identify the best response strategies for each player and prove the existence of Nash Equilibria in both games. Very recently, Zhu and Rass [18] propose a general framework that divides a general APT into three major temporal phases, and fits an individual game model to each phase, connecting the games at the transition points between the phases.

Bhatt et al. [19] propose a framework that models multi-stage attacks. Their intuition is to model behaviors using an Intrusion Kill-Chain attack model and defense patterns. The implementation of their framework is made by using Apache Hadoop. In a similar way, Giura et al. [20] use a large-scale distributed computing framework, such as MapReduce to consider all possible events coming from the monitoring process and process all possible contexts where the attack can take place.

Cyber Attack Attribution. In 2003, Wheeler published techniques for cyber attack attribution [21] while in 2008, Hunker et al. [22] highlighted the importance of cyber attack attribution and they present its challenges. Bou-Harb et al. [23] proposed an architecture that provides insights and inferences that help in attribution. Their architecture investigates attacks against cyber-physical systems. Qamar et al. [24] proposed a methodology that creates groups of threats based on their similarities in order to aid decision making; they also use ontologies. The importance and timeliness of enhanced cyber attribution is also pronounced by the fact that in 2016 USA Defense Advanced Research Projects Agency (DARPA) created a call which aims for the identification of the malicious parties responsible for the cyber attacks [25]. In the same year, the US Department of Defense awarded Georgia Institute of Technology a large research contract, to enable the development of the capability to quickly, objectively and positively identify the virtual actors responsible for cyber attacks [26].

DARPA [27] splits the attribution process in three distinct phases which run in parallel. First is the "Activity Tracking and Summarization" where ground truth is being formed through the collection of information from multiple sources (e.g., Ops desktop, mobile phone, IoT, captured C2 nodes, network infrastructure). Second phase is "Data Fusion and Activity Prediction" where some points of interest are picked from the previous phase and predictive profiles are being developed and ambiguous data associations are being captured across diverse data set. Third phase is "Validation & Enrichment" where adversary mistakes and externally observable indicators (e.g., Open-source intelligence, Commercial Threat Feeds, Network IDS/analytics etc.) are being identified.

NEON's novelty. The NEON framework takes into consideration all other approaches of APT detection proposed in the literature but it also introduces mechanisms that cope with rising challenges. NEON follows the DARPA guideline [27] to first build a ground truth of APT campaigns' related data through intelligence gathering. Then, it correlates collected data and uses honeypots to create points of interest that will allow the detection of zero-day exploits and novel attack techniques. During that process, it takes into consideration the challenge of adversarial machine learning to ensure that optimal decisions are taken, even in the presence of attacks aiming at disrupting defences that us machine learning. NEON also uses game theory to help in the arms race between the attacker and the defender by devising optimal defending strategies. Its final step is not only to pronounce the situational awareness, but also to contribute to the attribution of attackers.

4 Proposed Framework

In this section we discuss the various components of NEON, which are also illustrated in Fig. 2.

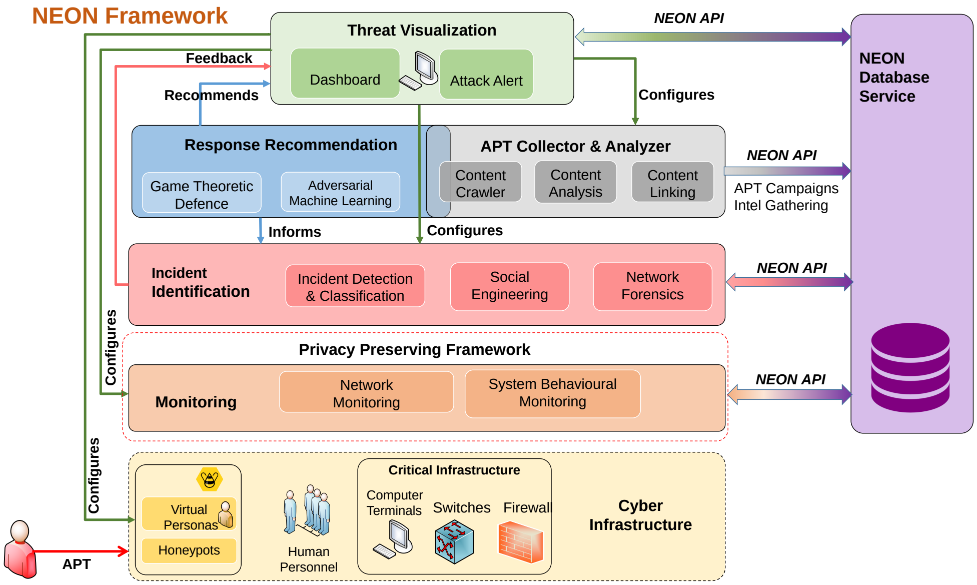

Fig. 2: NEON architecture.

4.1 APT Collector & Analyzer

To the best of our knowledge NEON APT Collector & Analyzer is the first of its kind to collect, sanitize and link different APT reports along with publicly available information on social media. There are a lot of APT repositories and crawlers which usually extract segments of the content of each report. The APT Collector & Analyzer is implemented by the following components [28].

Content Crawler. The various APT reports published by industry and academia are written in various ways making crawling a challenge. The NEON Content Crawler (CC) monitors a number of crawling points corresponding to repositories and social media sources with diverse content in an automatic and continuous fashion, with the goal to discover and collect APT reports and any other information related to them. CC is based on a crawling infrastructure capable of selectively collecting and scraping content related to APTs from Web resources by estimating (e.g., by using a distance function) their relevance to the topic of interest. This is based on an adaptive approach [29], which employs a semi-supervised methodology using unlabelled data within a supervised learning framework [30]. The crawling process is carried out using open source tools, such as Apache Nutch.

Content Analysis. The gathered APT related information is delivered to the NEON Content Analysis (CA) component for further processing. This extracts the named entities and concepts of interest (e.g., names of malware groups, APT names, temporal expressions, number expressions, domains used and their respective IP addresses) from the collected APT reports and from the other social media related sources. The collected information then subjects to linguistic analysis and is processed through a series of steps employing a pipeline of tools in the following order: sentence-breaking, tokenization, lexicons, part of speech tagging, text normalization, and eventually parsing [31]. Experiments are performed (i) with supervised/semi-supervised techniques and active learning techniques for parsing into shallow semantic structures as well as, (ii) with a dependency parser to generate deep syntactic structures. NEON then utilizes established linguistic approaches to named entity recognition (NER) and further enhances them by implementing machine learning techniques [32]. Furthermore, NEON integrates linguistic dependencies information, searches for the semantic types of relations used for the identification of candidate senses, and checks the overall semantic consistency of the resulting disambiguate structures.

Content Linking. Many different organizations, companies and research labs have analyzed the same APT campaigns. However, each research team gives a different name to the same campaign. In addition, the social media collected data is not directly linked to APT campaigns. The NEON Content Linking (CL) component solves this issue by combining the extracted APT concepts (e.g., ip address, campaign name), a process that needs to be efficient and scalable to cope with the large data volume and the highly heterogeneous nature of the data structure. CL fuses several sources of information with the goal to create links between the APT related information retrieved from CA. Given an APT report as input and using its meta-data and extracted concepts as multiple modalities, the fusion of all available modalities is based on a semantic filtering stage [33]. This process filters out the non-relevant results in a progressive way starting from the dominant modality, i.e., the attribute/concept that has been proven most effective in uni-modal APT-to-APT comparison. Comparing the similarity among all pairs of objects for all modalities is not a scalable process and involves at least a quadratic computational complexity. Based on a vector representation of the query and the collection, NEON first retrieves the top k_m-results which are relevant to the query with respect to the m-th modality and computes the corresponding similarities. The fusion of all progressively obtained similarities is a graph-based process and it leads to a ranked list of retrieved results.

4.2 Monitoring

Network Monitoring. NEON support network forensic activities by using an efficient component to retrieve relevant data that allows the ulterior processing, detection and classification of anomalies. Many, if not most, security breaches in an organization are facilitated or conducted through a network. Monitoring the network traffic the primary mechanism for detecting attacks. At key network points, the Network Monitoring (NM) component collects data for each network flow and for each network packet individually, utilizing technologies of Deep Packet Inspection (DPI), towards detection and classification of anomalies and intrusions. NM sends information about suspicious network activities to the NEON Incident Detection & Classification (IDC) component, described later on.

System Behavioural Monitoring. The detection of usage patterns, based on the knowledge of the normal system behaviour, becomes prominent for the detection of anomalies with respect to the normal behaviour of users and devices and the network traffic [34]. The System Behavioural Monitoring (SBM) component inspects network traffic in real-time and assesses the behaviour of systems nodes and their deviations from "standard" behaviour. It is dynamic in nature, meaning that it can adapt to changing environments. For instance, the input to the network sensors can be any systems or network communications feature, making it possible to detect a wide range threats.

Furthermore, situational awareness techniques are used for a context-based detection of anomalies when a holistic view of the whole cyber infrastructure is required, including the knowledge of interactions among participants (either human or physical devices) and their inter-dependencies. SBM incorporates supervised and unsupervised machine learning algorithms that may operate as standalone modules or as an ensemble that produces more accurate results but also requires more computational resources. The unsupervised component is capable of detecting anomalies within a system up to a point where the number of anomalous nodes exceeds the number of normal behaviour nodes and raise alarms displaying the anomalous node details using the Attack Alert (AAlrt) component, described later on. The supervised module supports two algorithms: (i) Support Vector Machines and (ii) Logistic Regression, which are capable of working in an online setting and can be reconfigured in real-time based on feedback from the system, environment or operator.

4.3 Incident Identification

Social Engineering. The Social Engineering (SE) component consists of two elements; the Virtual Personas (VPs) and the Honeypots.

Virtual personas. A versatile set of VPs is utilized for making active and attractive to the cyber attackers [1]. Both genders are present while the most frequent identities of personas will be determined and used within honeypots, described later in this section. The creation of VPs is based on real-world employees and their daily routines. Each VP becomes a unique prey to APTs and it appears to work in a different part of a cyber infrastructure, having different privileges. VPs are being continuously attended and updated according to the cyber infrastructures real workload so as to maximize the resemblance along with the realism. VPs' goal is to attract attacks promptly acquiring the appropriate knowledge, which contribute towards the attribution of the attacker.

Honeypots. These act as beacons of interest for the malicious parties throughout their lifetime, as VPs will appear more unaware of security procedures than other employees offering themselves to be exploited. The malicious actions will be recorded and continuously feed the collection process. The real-world data collected from those attacks supports the existing background knowledge on APT campaigns when correlated with network traces and system behavioural data, as well as APT reports. Consequently, SE is able to collect information from new APT campaigns before their manifestation, thus minimizing their impact. SE calculates a set of thresholds that will trigger the defending organisation to involve a human operator in the honeypot process. After this point, live operation monitoring helps to extract more information about the attacker by introducing controlled human errors that will accelerate the manifestation of the attack.

Incident Detection & Classification. The Incident Detection & Classification (IDC) component provides the capabilities of a Security Information and Event Management (SIEM) solution with the advantage of being able to handle large volumes of data and raise security alerts. A suitable correlation among the different types of information is paramount for discovering ongoing incidents that may lead to a serious compromise of the cyber infrastructure. Network and system behavioural data, data from honeypots, virtual personas logs, data from past incidents and data stored in the NEON database have to be correlated for the *identification of current incidents* or *estimation of future incidents*. Machine learning techniques, such as clustering methodologies or decision trees, combined with usage behaviour patterns, are being used to predict potential malicious events threatening cyber infrastructures. IDC performs real-time collection and analysis of security events; prioritization, filtering and normalization of the data gathered from different sources; consolidation and correlation of the security events to carry out a risk assessment and generation of alarms by the NEON Attack Alert (AAlrt) component.

Network Forensics. To enable collection of necessary forensic information that can be used as legal evidence in court, towards the attribution of cyber attackers, the Network Forensics (NF) component leverages an investigation methodology and relevant network forensic tools to analyze the collected network traffic. Using the OSCAR (Obtain information, Strategize, Collect evidence, Analyze, Report) methodology [35] we ensure that necessary forensic information is collected and can be used as legal evidence. The collection of evidence is achieved through the NM and BM components. NF consist of the following parts: (i) *Obtain information*: gather general information about the incident itself (date and time of the incident, persons and systems involved, what initially happened, what actions have been taken since then, who's in charge, etc.) and the environment (company, organisation) where it took place, that usually changes over time and go or change positions while at the same time equipment is phased out or replaced, new equipment is being added and configurations are changed; (ii) *Strategize*: since network data is very volatile, NF prioritizes forensic data acquisition according to the volatility of the sources, the potential value to the investigation and the effort needed to obtain them; (iii) *Collect evidence*: to allow collection of evidence all actions taken and all systems accessed should be logged, while the log should be safely stored and should include time, source of the evidence, acquisition method and the investigator(s) involved. Two major sources of network evidence exist: (a) network traffic captures, (b) log files that can be either collected at the generating system or a central log host; (iv) *Analyze*: different tools are used to recover evidence material; and (v) *Report*: NF provides a detailed forensic report as the final product of any forensic investigation. The report can be read by non-experts and it is in accordance with general forensic principles.

4.4 Response Recommendation

Game Theoretic Defence. Conventional defences against APTs are often deployed in an ad-hoc manner. NEON aims to take into account the understanding of the attackers' goals and the objectives of the infrastructure under attack. It then utilizes the Game Theoretic Defence (GTD) component to propose optimal cybersecurity actions against the APT attacker. GTD is called when signs of the adversary are confirmed and mitigation must take place. GTD ensures that optimal defending strategies, in the form of security tasks (e.g., security configurations, manual human actions), are undertaken. The response includes both the set of controls that are used to mitigate the attack actions as well as the way the tasks of a system administrator are prioritized to maximize their efficiency. GTD is based on the representation of the system under-attack in the form of a graph with different states, including both exploited and recovery states. GTD is based on a zero-sum game between the defender, which is the organisation that NEON protects and the APT attacker. The defender chooses among different cybersecurity portfolios and the attackers have a set of targets to exploit in system/network they have gained access to [36], [37], [38].

Adversarial Machine Learning. Adversarial machine learning (AML) is the study of robust machine (or statistical) learning techniques to an adversarial opponent, who aims to disrupt the learning (causative attacks) or the classification (exploratory attacks) and hence any subsequent decision making process with malicious intent [2]. For example, AML can make innocent data input to be classified as malicious and vice versa. NEON uses AML defences to prevent erroneous behaviour of security-related classifiers. In this way, NEON guarantees data trustworthiness thus increasing trust to the systems involved in undertaking cyber security related actions, such as intrusion detection. For example, NEON shall compute “optimal” thresholds for retraining a classifier as a result of a concept drift [39]. A first set of experiments have been undertaken as part of [40] to assess the performance of various classifiers in presence of adversarial samples of varied volume.

4.5 Threat Visualization

Dashboard. Given the heterogeneity and complexity of any data related to APTs, the visual analysis is done through a simple intuitive interface allowing for effective representation of the identified data patterns. NEON employs an interactive user-friendly visualization dashboard displaying real-time information acquired from NM, SBM, IDC, and GTD components, as well as from the APT Campaign Database. This is done by the NEON Dashboard (DsB) component. This offers visual analytics with several security-oriented data transformations and representations, including but not limited to network intrusion graphs, traffic histographs, temporal charts, location maps, and 3D visualizations, in an effort to simplify the highly complex data and provide a meaningful threat analysis. The main goal is to provide a highly customizable environment for users, attempting to balance between automation and control. DsB is built upon a tier-based architecture, where the higher-level tiers present a general overview of the data and the lower-level tiers display more detailed representations, allowing users to pull up information and drill down into specific details when needed. Finally, the graphical user interfaces built for the configuration of the NEON components provides users with the opportunity to handle and manage the operation of the NEON framework.

Attack Alert. Incidents threatening a cyber infrastructure can affect the system in different ways. Certain threats may not have great impact to the system when they compromise devices with no connections to the critical assets. Other threats may have greater impact on critical systems that, depending on the type of infrastructure, can even impact human lives (i.e., temperature sensors in nuclear plants). The Attack Alert (AAlrt) component interfaces with IDC to provide system administrators with localized and situated notifications. AAlrt delivers visual and audio notifications to the users through the dashboard aiming at increasing the user understanding and situational awareness. The alerts refer to the potential infection risks when a suspicious activity is detected by IDC.

As the component uses the results of classification to alert administrators about various current and future incidents, it allows prioritizing potential reactions against threats, depending on the foreseen effect in the infrastructure. To this end, assessment techniques are carried out in order to estimate the risk associated to a threat. Aspects such as the criticality of the system affected by a threat and the cost (either monetary or in term of resources consumption) of dealing with it or the speed of “threat propagation” across the system, may determine the appropriateness of mitigating certain threats in an effort to allocate limited resources in an optimal way. All these criteria are used to classify incidents, giving system administrators valuable information for an efficient and trustworthy management of cyber infrastructures.

5 Healthcare Use Case of NEON

Every established national healthcare system handles medical data and other sensitive data (insurance, payment, etc.). Additionally, it possesses a lot of other mechanisms which help the treatment process and in case a minimum delay appears in the process, it may even result in human casualties. In the chosen scenario, illustrated in Fig. 3, a healthcare system is being attacked. The attack takes place through phishing using social media and valuable information from a malicious insider. The latter provides the attacker with necessary information (**1. Gives Information**) to approach a real person, thus ignoring the honey-farm establishment. Because the specific person has been recruited recently and has not taken the training session, falls for the phishing attack (**2. Phishing Attack**). Consequently, the newly recruited employee clicks the malicious file sent through social media, which installs a binary that takes advantage of a zero-day vulnerability (**3. Executes**). It then hides itself by attaching its executable to a legitimate process in the deployed system. The attacker's final goal is to infect the whole healthcare system and break the electronic health service.

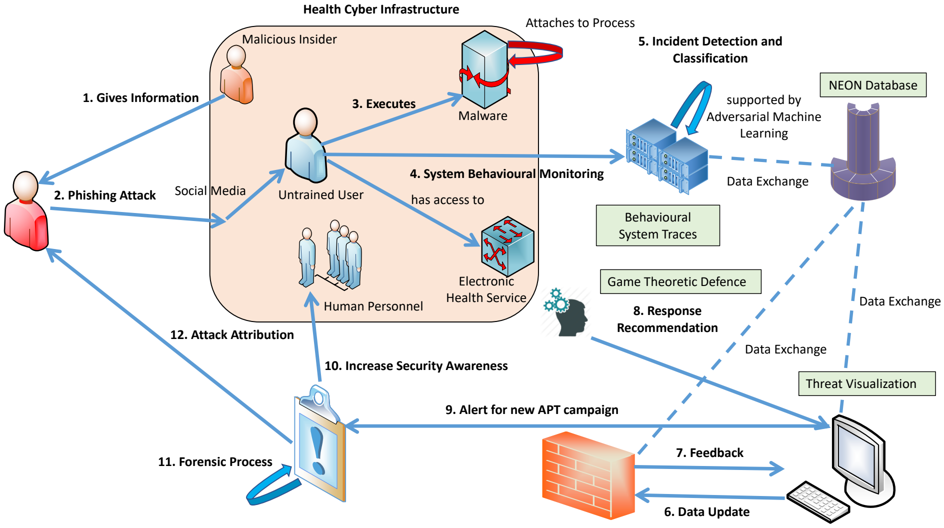

Fig. 3: Healthcare use case of NEON.

The attacker has managed to overcome NEON's implemented honey-farm establishment due to the help of the malicious insider. However, the execution of the malware from the untrained employee will produce system behavioural data (**4. System Behavioural Monitoring**). The recorded data populates the existing APT-related database, which stores data collected from APT reports and social media sources handled by the APT Collector & Analyzer components (CC, CA, CL) and the Network Monitoring component. All this data will be correlated with former known attacks (**5. Incident Detection and Classification**). The detection is supported by the Adversarial Machine Learning (AdvML) component as the adversary may use evasion techniques. As soon as the NEON database is updated with new APT data (**6. Data Update**), it will be possible to identify parts of known attacks in APT campaigns enabling their attribution and triggering attack alerts (AAlrt component) (**7. Feedback)**. The Game Theoretic Defence component (**8. Response Recommendation**) will propose optimal security responses. Consequently, any well-hidden or hibernated mechanism will be revealed and generated alerts (**9. Alert for known APT campaign**) will notify organization employees and increase their security awareness about the new threat (**10. Increase Security Awareness**). After the infrastructure has gained resistance against the APT campaign and it has collected evidence that includes network traces, system behavioural patterns, IPs and domain names, the Network Forensics component (**11. Forensic Process**), will deliver a report that is factual and defensible in detail in a court of law, in order for the law prosecution process to be initiated (**12. Attack Attribution**).

6 Conclusions

Enhanced attack attribution frameworks are in their infancy. At the same time APT becomes the most prominent threat paradigm. To address challenges that emerge from the above, this paper proposes the NEON framework. Its primary target is the collection and representation of intelligence about APT campaigns and then the correlation with monitoring activities. In NEON, honeypots with the help of virtual personas improve the detection capabilities of zero-day exploits and social engineering attacks. Game theoretic defences are incorporated into NEON to mitigate the actions of sophisticated APT attackers. Furthermore, adversarial machine learning supports data trustworthiness thus facilitating accurate APT detection and attribution and a threat management console visualizes and pronounces the situational awareness of people and critical infrastructures in NEON. Finally, network forensics generate evidence that lead to the attribution of malicious parties, which is the overall aim of NEON.

As future work, we aim to develop NEON for various use cases based on existing software tools and novel methodologies of partners. Given the complexity of the APT detection and attribution landscape, we envisage this to be a challenging task. Our plan is to develop the individual NEON components in the following order: (i) APT Collector & Analyzer, (ii) Monitoring, (iii) Incident Identification, (iii) Response Recommendation, and (iv) Threat Visualization.

References

1. Farinholt, B., Rezaeirad, M., Pearce, P., Dharmdasani, H., Yin, H., Le Blond, S., McCoy, D., Levchenko, K.: To catch a ratter: Monitoring the behavior of amateur darkcomet rat operators in the wild. In: IEEE Symposium on Security and Privacy, Ieee (2017) 770–787
2. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.: Adversarial machine learning. In: 4th ACM Workshop on Security and Artificial Intelligence, ACM (2011) 43–58
3. Pfleeger, S.L., Sasse, M.A., Furnham, A.: From weakest link to security hero: Transforming staff security behavior. Journal of Homeland Security and Emergency Management **11**(4) (2014) 489–510

4. Langner, R.: Stuxnet: Dissecting a cyberwarfare weapon. IEEE Security & Privacy **9**(3) (2011) 49–51
5. Kaspersky: Targeted cyber attacks logbook. https://apt.securelist.com/ (visited on 09-02-2018)
6. Symantec: Advanced persistent threats: A symantec perspective. https://www.symantec.com/content/en/us/enterprise/white_papers/b-advanced_persistent_threats_WP_21215957.en-us.pdf (visited on 09-02-2018)
7. ITU: Targeted attack trends. https://www.itu.int/en/ITU-D/Cybersecurity/Documents/2H_2013_Targeted_Attack_Campaign_Report.pdf (visited on 09-02-2018)
8. King, S.: Apt (advanced persistent threat) – what you need to know. https://www.netswitch.net/apt-advanced-persistent-threat-what-you-need-to-know/ (visited on 09-02-2018)
9. Cavelty, M.D.: Cyber-security and threat politics: US efforts to secure the information age. Routledge (2007)
10. Choo, K.K.R.: The cyber threat landscape: Challenges and future research directions. Computers & Security **30**(8) (2011) 719–731
11. Giura, P., Wang, W.: A context-based detection framework for advanced persistent threats. In: International Conference on Cyber Security, IEEE (2012) 69–74
12. Virvilis, N., Gritzalis, D.: The big four-what we did wrong in advanced persistent threat detection? In: 8th International Conference on Availability, Reliability and Security, IEEE (2013) 248–254
13. Jasek, R., Kolarik, M., Vymola, T.: Apt detection system using honeypots. In: 13th International Conference on Applied Informatics and Communications. (2013) 25–29
14. Chen, P., Desmet, L., Huygens, C.: A study on advanced persistent threats. In: IFIP International Conference on Communications and Multimedia Security, Springer (2014) 63–72
15. Friedberg, I., Skopik, F., Settanni, G., Fiedler, R.: Combating advanced persistent threats: From network event correlation to incident detection. Computers & Security **48** (2015) 35–57
16. Marchetti, M., Pierazzi, F., Colajanni, M., Guido, A.: Analysis of high volumes of network traffic for advanced persistent threat detection. Computer Networks **109** (2016) 127–141
17. Hu, P., Li, H., Fu, H., Cansever, D., Mohapatra, P.: Dynamic defense strategy against advanced persistent threat with insiders. In: IEEE Conference on Computer Communications, IEEE (2015) 747–755
18. Zhu, Q., Rass, S.: On multi-phase and multi-stage game-theoretic modeling of advanced persistent threats. IEEE Access **6** (2018) 13958–13971
19. Bhatt, P., Yano, E.T., Gustavsson, P.: Towards a framework to detect multi-stage advanced persistent threats attacks. In: Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on, IEEE (2014) 390–395
20. Giura, P., Wang, W.: Using large scale distributed computing to unveil advanced persistent threats. Science Journal **1**(3) (2012) 93–105
21. Wheeler, D.A., Larsen, G.N.: Techniques for cyber attack attribution. Technical report, Institute for Defense Analyses Alexandria VA (2003)
22. Hunker, J., Hutchinson, B., Margulies, J.: Role and challenges for sufficient cyber-attack attribution. Institute for Information Infrastructure Protection (2008) 5–10
23. Bou-Harb, E., Lucia, W., Forti, N., Weerakkody, S., Ghani, N., Sinopoli, B.: Cyber meets control: A novel federated approach for resilient cps leveraging real cyber threat intelligence. IEEE Communications Magazine **55**(5) (2017) 198–204

24. Qamar, S., Anwar, Z., Rahman, M.A., Al-Shaer, E., Chu, B.T.: Data-driven analytics for cyber-threat intelligence and information sharing. Computers & Security **67** (2017) 35–58
25. DARPA: Enhanced attribution federal project. https://govtribe.com/project/enhanced-attribution (visited on 09-02-2018)
26. Kintis, P., Miramirkhani, N., Lever, C., Chen, Y., Romero-Gómez, R., Pitropakis, N., Nikiforakis, N., Antonakakis, M.: Hiding in plain sight: A longitudinal study of combosquatting abuse. In: ACM Conference on Computer and Communications Security, ACM (2017) 569–586
27. Keromytis, A.: Enhanced attribution. https://www.enisa.europa.eu/events/cti-eu-event/cti-eu-event-presentations/enhanced-attribution/ (visited on 09-02-2018)
28. David Westcott, K.B.: Aptnotes. https://github.com/aptnotes/data (visited on 09-02-2018)
29. Meusel, R., Mika, P., Blanco, R.: Focused crawling for structured data. In: 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM (2014) 1039–1048
30. Triguero, I., García, S., Herrera, F.: Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowledge and Information Systems **42**(2) (2015) 245–284
31. Olston, C., Najork, M.: Web crawling. Foundations and Trends in Information Retrieval **4**(3) (2010) 175–246
32. Cimiano, P.: Ontology learning from text. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications (2006) 19–34
33. Gialampoukidis, I., Moumtzidou, A., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Retrieval of multimedia objects by fusing multiple modalities. In: ACM on International Conference on Multimedia Retrieval, ACM (2016) 359–362
34. Pitropakis, N., Pikrakis, A., Lambrinoudakis, C.: Behaviour reflects personality: detecting co-residence attacks on xen-based cloud environments. International Journal of Information Security **14**(4) (2015) 299–305
35. Davidoff, S., Ham, J.: Network forensics: tracking hackers through cyberspace. Volume 2014. Prentice hall Upper Saddle River (2012)
36. Fielder, A., Panaousis, E., Malacaria, P., Hankin, C., Smeraldi, F.: Decision support approaches for cyber security investment. Decision Support Systems **86** (2016) 13–23
37. Fielder, A., Panaousis, E., Malacaria, P., Hankin, C., Smeraldi, F.: Game theory meets information security management. In: IFIP International Information Security Conference, Springer (2014) 15–29
38. Fielder, A., Konig, S., Panaousis, E., Schauer, S., Rass, S.: Uncertainty in cyber security investments. arXiv preprint arXiv:1712.05893 (2017)
39. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine learning **23**(1) (1996) 69–101
40. Nikhi, B., Giannetsos, T., Panaousis, E., Took, C.C.: Unsupervised learning for trustworthy IoT. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (2018)