Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Avviso pubblico per la presentazione di Proposte di intervento per la realizzazione di attività di ricerca fondamentale relative al Partenariato Esteso SERICS (PE00000014), nell'ambito dello Spoke 3 "Attacks and Defences" (UNIVERSITA' DEGLI STUDI DI CAGLIARI) ammesso a finanziamento con Avviso Pubblico nr 341 del 15-02-2022 "Partenariati estesi alle università, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base" – nell'ambito del Piano Nazionale di Ripresa e Resilienza, Missione 4 "Istruzione e ricerca" – Componente 2 "Dalla ricerca all'impresa" – Investimento 1.3, finanziato dall'Unione europea – NextGenerationEU – CUP: F53C22000740007

Allegato Tecnico

The "Attack and Defense" Spoke aims to analyze emerging attack methodologies and develop advanced systems for their detection and for the identification of guidelines for the design of computer systems characterized by reduced vulnerabilities with respect to new attack categories. The detailed objectives can be divided into four macro categories: i) Development of advanced tools for malware analysis and software analysis aimed at identifying vulnerabilities that could be exploited by malware; ii) Development of network traffic analysis tools to identify communications related to ongoing attacks; iii) Development of machine learning systems that are robust to attacks and from which it is possible to extract knowledge aimed at creating more advanced tools for the timely analysis of attacks and their early detection; iv) Analysis of the "human factors" involved in an attack with the development of tools for the analysis and correlation of information from OSINT sources and for the defense and prevention of attacks based on social engineering techniques.

This Spoke aims to achieve its objectives within four complementary projects. These projects address core research issues for developing effective defenses against sophisticated attacks. The **CSS** project will address social engineering and human factor issues that account for the preliminary stages of preparation and execution of a cyber attack. In the later stages of an attack, sophisticated techniques are used to evade detection and response mechanisms. The **COVERT** project will develop analysis and detection tools for the early detection of silent software vulnerabilities and obfuscated malware. The detection of sophisticated malicious activities through **network** traffic analysis requires the development of advanced tools addressed within the GERONIMO project. Attack detection techniques increasingly use machine learning methods whose security is essential to consider them reliable. The **SOS AI** project will focus on attacks against machine learning systems and will develop appropriate design techniques for the execution of these algorithms in a hostile environment. The four projects will work synergistically by sharing data, case studies, and results.

This Spoke aims to attain a twofold goal: on the one hand, to increase the capabilities for the independent analysis of sophisticated attacks, that is, attacks designed to evade existing detection techniques: on the other hand, to develop advanced capabilities for the design of information and defense systems that can be resilient to the continuous change and sophistication of cyber-attacks, especially those targeting critical infrastructures.

The document is organized as follows. It first outlines the overall plan of the activities and milestones of Spoke 3. Then, it provides useful details about the projects included in Spoke 3. Finally, the section "Open Calls / Bando a Cascata" details the objectives of the tasks of each project that are the subject of this notice.

Plan of activities and milestones

The project started on 1st Jan 2023 (M1). The duration of the project is 36 months.

The overall plan of the activities and milestones of the Spoke is summarized in the diagram shown in Figure 1. The figure details each type of activity, separated by horizontal bars, to which the partner exposes the costs of the project. The figure also displays the checkpoints at which the Spoke leader and partners must summarize the findings obtained in the corresponding period as vertical red lines. Analogously, after the selection of the most-suited proposal for the Open Call corresponding to each project, the winning candidate should provide technical reports, one at each checkpoint in the diagram (red lines). The technical report will describe the findings obtained in the corresponding period and a software implementation of the best-performing techniques and best-suited method.

In addition, the winning candidate should monthly provide an update on the activities carried out on the ongoing Open Call.

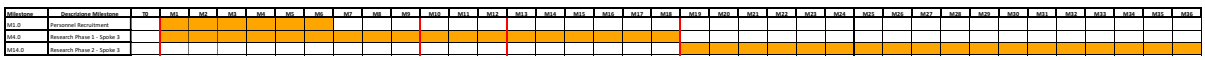


Figure 1. Gantt diagram

Milestones are set at the end of each period, when the results obtained in the last considered ones will be revised, and the work to be done in the subsequent period will be planned in light of the results obtained so far.

For each project scope (i.e., Ambito Progettuale) included in Spoke 3, the sections below provide a detailed breakdown of the tasks and their corresponding main objectives. The selected proposal must successfully complete the tasks outlined below, ensuring that the requirements and objectives of the project milestones are met.

Each proposal should focus on addressing a specific project scope. In the case of more than one partner participating in the same proposal for addressing the same project scope, each of them must clearly state their role, expected outcomes, and corresponding budget.

Additionally, the proposal has to plan the activities over time by producing a GANTT chart including milestones in accordance with the overall project GANTT reported in Figure 1 and respect the deadlines for documentation and software deliveries.

CSS

Cyber Social Security

Abstract

It is a common viewpoint that the combination of data coming from social media, smart-phones and from urban sensors can actually enable the ability to carry out in-depth analyzes and understand complex phenomena based on human behavior, opening new scenarios for the development of numerous innovative services and applications. By following this research line, the recent paradigm of Social Sensing further emphasized this vision, since it proposed an integrated model in which users themselves are turned into sensors, entities that produce simple rough information which is processed and aggregated in order to generate some valuable human-based findings obtained through the combination and merge of individual-based data.

Beyond sensing applications, as those focusing on tracking vehicles to avoid traffic congestions or healthcare tracking and predicting people's lifestyle, a big research effort has been made to analyze text-based signals, such as those coming from social networks like Twitter or Facebook. The reason is twofold: first, methodologies for Natural Language Processing (NLP) rely on very consolidated and effective algorithms, thus it is relatively simpler to process textual data rather than audio, video or especially environmental-based ones. Second, despite its size grows more slowly than video or audio data, textual content represents a very rich, interesting and valuable information source. Furthermore, in a scenario where cyberspace events impact on the real world and influence the political, social and cultural spheres, it is essential to have the cognitive, methodological superstructures as well as the cyber-physical infrastructures necessary to guarantee the resilience of civil society. The purpose of this project is to affect these issues through the proposition of multidisciplinary methods, techniques and tools (IT; psychological, economic, legal, engineering, related to social sciences) capable of operating a Cyber-Social risk management in civil society. To this end, it is necessary to reinterpret the functions of Cyber Security in Cyber Social contexts:

• Detection: characterize, identify, understand and predict significant cyber-mediated events and changes in human, social, cultural and political behavior as well as the methods for monitoring and protecting "social" end-points, thus being able to operate with devices (IT and IoT) and diversified information sources (OSINT/CLOSINT) taking into account the national and international legal framework (GDPR; NIS; CyberSecurity Act).

• Prevention: redefine the processes of census and prevention of "accidents" in the light of new critical assets (individuals, groups, communities, software applications and infrastructures for the public service, etc.), including elements of physical, organizational and applicative security as well as socio-political, economic, psychological and legal context.

• Response: defining intervention and cooperation protocols between the main players in civil society in order to guarantee resilience and social security, including through homeland security technologies and the fight against cyber terrorism and cybercrime.

The review of the Detection-Response-Prevention cycle will also clarify the limits within which it is possible to find and manage information while protecting the citizen's right to privacy and the security of civil society.

Work Breakdown Structure

WP1 - Innovations for Cyber Social Detection

Characterize, identify, understand and predict significant cyber-mediated events and changes in human, social, cultural and political behavior as well as the methods for monitoring and protecting "social" end-points, thus being able to operate with devices (IT and IoT) and diverse information sources (OSINT/CLOSINT) taking into account the national and international legal framework (GDPR; NIS; CyberSecurity Act) and geo-polical scenario.Task

- Task 1.1 Methods for extraction of social sensor data
- Task 1.2 Methods for extraction of urban sensors data
- Task 1.3 Demonstrators - software and Infrastructure for extraction of cyber social data

WP2 - Innovations for Cyber Social Response

Defining intervention and cooperation protocols between the main players in civil society in order to guarantee resilience and social security, including through homeland security technologies and the fight against cyber terrorism and cybercrime.

- Task 2.1 Methods for processing of social streams
- Task 2.2 Methods for processing of urban stream
- Task 2.3 Definition of CSS Response process patterns
- Task 2.4 Demonstrators - software and Infrastructure for analyzing and processing urban and social streams

WP3 - Innovations for Cyber Social Prevention

Redefine controls and processes for the prevention of "accidents" in the light of new critical assets (individuals, groups, communities, software applications and infrastructures for the public service, etc.), including elements of physical, organizational and applicative security as well as socio-political, economic, psychological and legal context

- Task 3.1 Prevention controls and functions for managing CSS
- Task 3.2 Guidelines for preserving rights, privacy and security in the CSS context

Open Calls / Bando a Cascata

Project Scope: CSS

WP1 Task 1.3 - Demonstrators - software and Infrastructure for extraction of cyber social data

The purpose of the task is the development, through the integration of existing solutions (hw/sw) and ad hoc developed code, of demonstrators useful for detecting and automatically collecting data and information for cyber social security.

Main Objectives

- Provide connectors capable of collecting data streams from urban and social sensors.
- Provide a high-level interface for Data Visualization useful for supporting cyber-social analyst in interpreting the data collected.

WP2 Task 2.4 - Definition of CSS Response process patterns.

The purpose of the task is to define a set of process patterns to be composed ad used, in consideration of the social and urban context, for structuring and executing effective response plan in order to address social and urban resilience.

Main Objectives

- Define a methodology for the definition of process patterns useful for managing the response phases in Cyber-social contexts.
- Define the structure of a knowledge base for the collection of process patterns
- create a set of ready-to-use process patterns for managing the response phase in the main reference contexts (urban and social)

WP2 Task 2.4 - Demonstrators - software and Infrastructure for analyzing and processing urban and social streams.

The purpose of the task is the development, through the integration of existing solutions (hw/sw) and ad hoc developed code, of demonstrators useful for processing cyber social streams.

Main Objectives

- Produce an integrated hw/sw platform for the automatic analysis of social and urban streams through the use of NLP and AI algorithms.
- Model/import defined process patterns and enact them through a BPM engine for providing a response to cyber-social critical events.

WP3 Task 3.2 - Guidelines for preserving rights, privacy and security in the CSS context

The purpose of the task is the definition, after having carried out a survey of the state of the art, of guidelines for preserving rights, privacy and security in the CSS context. These guidelines should therefore be easy to use and should guide the design of demonstrators and the activity of Cyber Social Security operators.

Main Objectives

- Provide a state of art analysis of the current regulation for preserving rights, privacy and security in cyber-social contexts
- Provide ready-to-use guidelines to guide data collection and analysis in cyber-social contexts in compliance with privacy and current legislation

COVERT

In searCh Of eVidence of stEalth cybeR Threats

Abstract

An increasing percentage of cyber attacks are characterized by a chain of different components whose harmful effect is realized only in the presence of specific contexts such as configurations, vulnerabilities of software components, etc. These attacks, known as Advanced Persistent Threats (APT), are characterized by the execution of malware "invisible" to detection systems, and often exploit vulnerabilities that remain silent for a long time, and have an extremely slow activation. The difficulty in identifying attacks is due on the one hand to execution profiles that are similar to those exhibited by legitimate activities, and, on the other hand, to the use of specific components designed to obfuscate the code or evade or mislead the analysis and detection tools. The growing sophistication of attacks makes their detection extremely difficult, requiring the development of new models for the analysis and detection of attacks and new methodologies that provide for the cooperation and integration of several different approaches.

The partners involved in this project will work jointly in producing a Threat Model aimed at relating the techniques and tools used in modern attacks with the malicious goals of the attackers. While in the past it was possible to clearly associate specific code fragments of program behaviour to a clear malicious intent, advanced attacks distribute the code and activities leading to the malicious goal in space and time, so that the analysis and detection tasks need to be reformulated. The complementary expertise of each partner will provide the unique contribution to the definition of an effective threat model. In the following, the specific methodologies that will be developed within this project are detailed.

To this end, we will study and develop advanced software and network traffic analysis methodologies and, in particular, tools for the early detection of vulnerabilities, and the timely detection of harmful components. This goal will be attained through the integration of complementary models and techniques for static and dynamic analysis supported by artificial intelligence and machine learning techniques as well as information obtained through OSINT.

Aspects related to the collection of large amounts of data both at the network level, and from software repositories and execution traces of binaries will also be addressed.

Thanks to the collaboration with companies, tests on different case studies will be carried out and proofs of concepts will be released to make the results available to the research community.

Work Breakdown Structure

WP1 - Analysis and detection of Evasive Malware and Stealth attacks

WP Description: The evolution of techniques for threat analysis and detection forced malware developers to increase the defense mechanisms embedded in their attacks. Those defense mechanisms comes in the form of evasion techniques and stealth behavior. Modern threats can include evasion techniques that are able to detect if they are executed in a sandbox whose goal is to detect potential harmful behavior. For example, when an evasive malware sample detects the execution in a sandbox, then it exhibits a non-malicious behavior, so that so that they are flagged as not harmful, and their execution is allowed in a real environment. Threats can also exhibit an execution profile similar to that of legitimate applications, so that it is difficult to distinguish it without creating a large number of false alarms. The code itself might include different obfuscation techniques that prevent a thorough static analysis of the code.

The goal of this WP is to analyze the evasion and obfuscation techniques employed by modern malware both in traditional execution environments, and in industrial control systems, and deploy novel solutions that allows the analysis and detection of sophisticated malware samples.

Task 1.1 - Static and dynamic analysis frameworks to support evasion-detection techniques

Task 1.2 - Evasive and Stealth attacks in industrial control systems.

WP2 - Measuring the robustness of detection systems against evasion attacks

WP Description: Modern threats leverage on a large variety of weaknesses and vulnerabilities in operating systems, execution environments, communication protocols, and applications. Vulnerabilities often affect libraries that are part of applications, and whose correction is far

from being straightforward. Tracking the impact of vulnerabilities, and predicting new vulnerabilities is a new challenge that needs to be addressed to be prepared in the deployment of defence in depth approaches. In particular, most recent vulnerabilities allow the development of evasive and stealth malware. Thus, their prediction allows the design and development of advance threat detection mechanisms that can face novel stealth and evasive behavior. Accordingly, this WP aims at developing methodologies and tools for the analysis of known vulnerabilities in order to develop approaches allowing to predict the likelihood of new vulnerabilities and their impact on detection mechanisms. This WP will also develop techniques for the categorization of evasive and stealth techniques with the aim of developing a mutation engine to measure the impact of new evasion and stealth techniques on detection systems. While the above two objectives will be based on the analysis of data from public repositories of weaknesses and vulnerabilities, a third objective of this WP is to develop a formal framework for measuring the impact of evasive and stealth techniques employed by modern threats on detection systems.

Task 2.1 - Large-scale repository mining for vulnerability analysis and impact prediction

Task 2.2 - Definition of a taxonomy of anti-analysis/evasion techniques and design of a mutation engine to generate evasive samples for testing existing analysis environments

Task 2.3 - Development of a formal framework to model and measure the effect of evasion techniques on the precision of static and dynamic analysis.

WP3 - Machine Learning Approaches to Obfuscated and Evasive Attack Analysis and Detection

WP Description: In the past decade Machine learning and deep learning approaches definitely moved from being the topic of a small area in the scientific literature on cybersecurity, to the mainstream of academic and industrial research. Nowadays, all cybersecurity products are based directly or indirectly on machine learning approaches, for their ability to generalize from examples that allows detecting never-before-seen attacks. Attackers are consequently improving their threats to evade detection through stealth and evasive techniques specifically developed to exploit the weaknesses of current machine learning solutions. This WP will focus on the development of machine learning approaches and techniques tailored to face the increased sophistication of threats in terms of their evasion capabilities, and the related need to

extract explanations that allows retrieving the knowledge gathered from data. Deep learning and advanced neural network approaches are currently being extensively evaluated as detectors, but they are showing to be extremely vulnerable to easy-to-implement evasion techniques. Thus, their use in cyberthreat detection should include specific training approaches to exploit the generalization capabilities without introducing vulnerabilities that can be exploited by evasion techniques. Malware detection, network intrusion detection and security information and event management systems will be the areas in which these approaches will be developed. Machine learning approaches will also be developed for the analysis of source code repositories for the early detection of vulnerabilities.

Task 3.1 - Development of context-based techniques for static and dynamic evasive threat detection inspired by deep learning approaches

Task 3.2 - Design and development Large Source Code Scanners for Large scale silent Vulnerability detection in Free and Open Source Software

Task 3.3 - Use of OSINT information to improve early detection capabilities of SIEM

Task 3.4 - Development of a Network-IDS based on Self-Organizing Incremental Neural Networks

Task 3.5 - Explainable AI techniques for enhanced malware analysis and improvement of detection techniques

Open Calls / Bando a Cascata

Project Scope: COVERT

WP2 - Measuring the robustness of detection systems against evasion attacks

Task 2.2 - Definition of a taxonomy of anti-analysis/evasion techniques and design of a mutation engine to generate evasive samples for testing existing analysis environments

Task Description: The aim is to catalog, characterize and systematize malware evasion techniques, trying to derive insight into future or unknown potential techniques (a "periodic table" approach, with an analogy). Additionally, a mutation engine will be designed to apply to malicious samples to test commercial and open-source sandbox and dynamic analysis system designs for resilience and improve them. Finally, based on the developed evasive technique taxonomy, the impact of evasion on modern antimalware and endpoint protection solutions will be tested.

Task 2.3 - Development of a formal framework to model and measure the effect of evasion techniques on the precision of static and dynamic analysis.

Task Description: To understand the impact of the increasing APT sophistication on program analysis, a formal framework for modeling and measuring the effects that program transformations have on static and dynamic analysis precision will be developed. This framework would allow to: (1) evaluate the effectiveness of threat detection methods in terms of the effects that commonly used evasion techniques have on their precision; (2) prioritize the analyses that need to be improved according to the impact that common evasion transformations have on their precision; (3) compare the resistance of detection methodologies concerning their loss of precision in the presence of evasion techniques. To this end, a general formal framework for program transformations to tune the precision of dynamic analysis and define a measure of precision/imprecision for static and dynamic analysis will be developed.

WP3 - Machine Learning Approaches to Obfuscated and Evasive Attack Analysis and Detection

Task 3.2 - Design and development Large Source Code Scanners for Large scale silent Vulnerability detection in Free and Open Source Software

Task Description: This task aims to cover the threat of direct attacks launched by malware and the threat to the supply chain. In this scenario, vulnerabilities are injected into the code during the development phase so that new and undetectable malware (as it exploits vulnerabilities created on purpose) can exploit them. The objective is to create a methodology that will allow practitioners to clearly understand the findings generated by vulnerability scanners for possibly silent vulnerabilities. This includes (1) the ability to precisely locate the alleged vulnerable code fragment (e.g., a fine-grained code fragment or even specific lines of code), (2) provide support for understanding whether such code fragment is actually affected by a suggested vulnerability.

Task 3.3 - Use of OSINT information to improve early detection capabilities of SIEM

Task Description: The use of OSINT data sources to improve SIEM capabilities will be studied with the goal of providing a real-time analysis of the generated security alerts. OSINT sources such as social network platforms, blogs and forums, news aggregators, darknets, cryptocurrencies transactions and platforms, Operation Technology honeypot information, gray literature, etc. will be considered. This task will produce the specification, analysis, design, and development of such a SIEM (including detection probes) using advanced AI algorithms.

Task 3.4 - Development of a Network-IDS based on Self-Organizing Incremental Neural Networks

Task Description: This task will design and develop a new network intrusion detection system with continuous learning based on Self-Organizing Incremental Neural Network (SOINN). In particular, the use of "bigraphs", particular graph structures capable of simultaneously representing both the position of the nodes in the network and their logical connections, will be investigated. Given this model, it is possible to formally specify the security policies that must

be implemented and maintained. The model is useful for statically verifying the correctness of given security policies with respect to the pre-set goals.

Task 3.5 - Explainable AI techniques for enhanced malware analysis and improvement of detection techniques

Task Description: Inner layers of deep classifiers, specifically designed to address malware detection, can capture significant medium-level features that discriminate between malicious and legitimate objects. Unfortunately, a learning approach that can transform low-level features into meaningful intermediate features associated with malicious behavior is currently missing. Explainable AI (xAI) approaches tailored to malware detection will be devised to extract the part of the input (e.g., network traffic, source code or binary code) that make the highest impact on the decision of a classifier. To this end, general xAI approaches such as feature attribution, explanation-by-example, counterfactuals, etc., will be investigated and tailored to the context of threat analysis and detection.

GERONIMO

GEneralized Real-time On-line National Internet MOnitoring Infrastructure

Abstract

The strategic role of network connectivity in many of the vital sectors of society made the Internet one of the critical infrastructures whose availability is crucial for any organization. Thus, there is a need for new Internet monitoring and surveillance solutions. Starting from these premises, the project focuses on a selective Internet surveillance solution, consisting of a distributed monitoring infrastructure operating on a national scale, able to dynamically collect, filter, classify and analyze traffic and trigger real-time alerts based on the detection of specific attack patterns. Selectivity must be provided by traffic analysis engines with advanced inferential capabilities empowered by AI/ML technologies and integrating honeypot capabilities to better detect hostile activities also at the application level. Detection effectiveness relies on the ability to correlate information gathered in distinct observation points. Correlation enabled by federated technologies will be extremely helpful to build new knowledge from the combination of multiple, apparently unrelated events, allowing the development of more sophisticated and reliable models that are able to better understand the deepest dynamics governing the traffic flows or malware activities of interest. This can be useful in developing automatic detection, alerting and filtering mechanisms effective against next-generation APTs characterized by polymorphic and adaptive behaviors and will foster synergies between research, LEAs, technology manufacturers and ISPs, for countering large-scale attacks.

Work Breakdown Structure

WP1 - Developing Selective Traffic Capture and Monitoring Tools

The WP will explore the design of new highly selective Internet surveillance systems, mainly based on Deep Packet Inspection and advanced protocol analysis technologies able to capture and store in a privacy-aware and secure way only the traffic data that may be of interest due to a specific behavior.

- Task 1.1 Hierarchical and adaptive traffic capture and analysis, collecting statistical features (UNISA + OPEN CALL 1)
- Task 1.2 Anonymization techniques for privacy-aware storage of traffic data (UNISA)
- Task 1.3 Digital Integrity of Multimedia Contents (UNIGE + UNISA + OPEN CALL 1)

WP2 - Traffic Classification and Anomaly Detection Technologies

The detection of threats and anomalies within the huge amount of normal communication flows relies on the ability to correlate local information gathered in multiple network nodes through intelligent detection and/or classification engines relying on dynamic knowledge bases as well as on traffic models able to recognize the behavior of traffic flows, spot outliers, and hence pick the classic haystack in the needle.

- Task 2.1 AI and ML-techniques for Internet traffic modeling (UNISA)
- Task 2.2 Developing classification and detection schemes (UNISA)
- Task 2.3 Assessing and Validating models and classifiers/detectors (UNISA)

WP3 - Honeypot modules and their coordination

Surveillance engines also integrate honeypot capabilities, complementing traffic sensors to better detect malware activities also at the application level. This WP will be devoted to the exploration of the honeynet architecture and malware analysis activities.

- Task 3.1 Honeypot architecture and deployment (UNISA + UNIGE).
- Task 3.2 Honeypot Coordination- Results Correlation (UNISA)
- Task 3.3 Advanced Malware detection facilities in honeypots (UNISA + OPEN CALL 2)

WP4 - Mitigation and Automatic Reaction

Once an attack/menace has been detected, automatic containment technologies, available on firewalls or border routers, such as packet and content filtering, can be used to block attack flows or hostile origins. This WP explores the reaction framework in the state-of-the-art Internet scenario.

- Task 4.1 Traffic Filtering Distribution Mechanisms (UNISA)
- Task 4.2 - Automatic Reaction Triggering - Decision Engine (UNISA+ OPEN CALL 3)

Open Calls / Bando a Cascata

Project Scope: GERONIMO

WP1 Task 1.1 - Hierarchical and adaptive traffic capture and analysis, collecting statistical features.

The main aim of this activity will be live traffic dynamic clustering and tagging, where per-node and per-flow statistical features have to be collected and analyzed in real time. Once a suspect behavior has been spotted, the depth of the analysis will be enhanced by adopting a hierarchical model. In addition, efficient representation and compression formats will be developed for both short and long-term storage of traffic data.

Main Objectives:

- Study and design of the selective capture facilities and of the automatic triggering mechanisms with associated models, attack representation formalisms and schemes.
- Development of a prototype implementation of the above facilities and mechanisms.
- Functional and performance evaluation and security assessment of the above facilities and mechanisms.

WP2 Task 1.2 - Anonymization techniques for privacy-aware storage of traffic data.

The effort required in this task essentially complements the development of new traffic interception solutions fully compliant with current privacy-enforcement regulations and forensics best practices. More precisely, the specific focus of this activity involves the combination of encryption algorithms and anonymization schemes, to collect network evidence for use in strategic surveillance as well as prevent and trace any illegal manipulation of the data and ensure their validity as legal proofs with law-enforcement authorities.

Main Objectives

- Study and design of anonymization/pseudonymization and selective encryption mechanisms for the captured traffic.
- Prototype implementation of the above mechanisms
- Security analysis, privacy assessment and performance evaluation of the above mechanisms.

WP3 Task 3.3 - Advanced Malware detection facilities in honeypots

Several research directions in static and dynamic malware analysis, empowered by intelligent, adaptive and self-learning detection engines will be explored for coping with polymorphic and metamorphic viruses and ransomware. Also, the effectiveness of compromise indicators, the application of formal methods to malware recognition, as well as the representation through images of the malware will be studied.

Main Objectives:

- Study end definition of the effectiveness of compromise indicators in honeypots to rank the effectiveness of attacks.
- Study and analysis of the application of formal methods to malware recognition.
- Design and Development of an interface/interworking mechanism between honeypots and external malware detection facilities.

WP4 Task 4.2 - Automatic Reaction Triggering - Decision Engine

The automation of the decision-making process related to triggering the proper traffic filtering actions based on specific alerts is the main goal of this task.

The main focus of the activity is the creation of a flexible and effective decision engine able to implement the association between attacks and reaction strategies by identifying the set of defense options and their configurations.

Main Objectives:

- Study and Design of a representation of specific attacks, associated defence options and reaction strategies (e.g., decision tables) and their mapping into countermeasures configuration (e.g., ACLs)
- Design and Development of a prototype of the decision engine and its interface to the traffic filter distribution mechanism (based on BGP Flowspec).
- Functional evaluation of the decision engine. This objective involves conducting rigorous experiments to assess the overall effectiveness of the implemented prototype, determine its limitations, and fix/refine them based on empirical evidence.

SOS AI

Science and engineering Of Security of Artificial Intelligence

PI: Fabio Roli (fabio.roli@unige.it)

Partners

UNIBA (University of Bari Aldo Moro), UNICA (University of Cagliari), UNIGE (University of Genova), SSSA (Scuola Superiore Sant'Anna of Pisa), UNIVE (Università Ca' Foscari Venezia), TIM (TIM – Telsy)

Abstract

Microsoft reported a dramatic increase of attacks on commercial systems based on Artificial Intelligence (AI) and machine learning (ML) algorithms over the past years. Notably, Microsoft pointed out that companies usually lack the knowledge and tools to secure their ML-powered systems. The Gartner's «2020 Top 10 Strategic Technology Trends» report predicted that 30% of future cyber-attacks might consist of attacks against the AI components of the computerized systems. Over the last ten years, the scientific community proposed many techniques to prevent the execution of attacks against AI/ML, or at least to detect them. However, in most of the cases, these attacks and defenses have been designed to work in laboratory conditions under simplified or unrealistic assumptions that do not consider the requirements of cybersecurity applications (e.g., the practical feasibility of the attacks is often not considered). Given this background situation, the SOS AI project wants to tackle the following 5 open challenges in an integrated way:

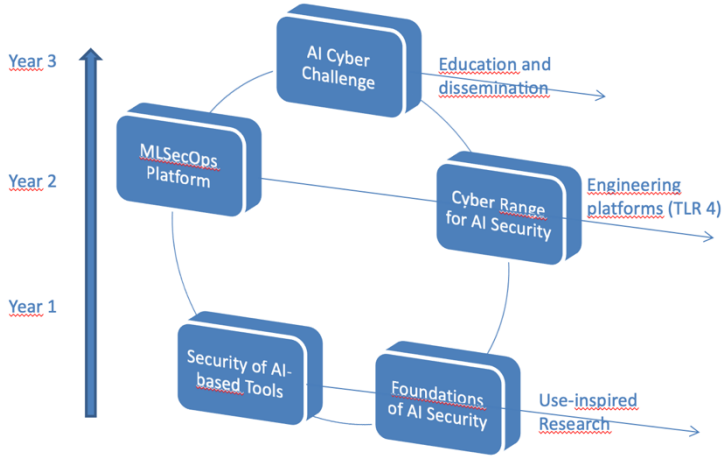
1. Foundations of AI security: the theoretical foundations of machine learning have not been originally thought considering intelligent and adaptive attackers who can manipulate input data to purposely subvert the learning process, that is exactly the case of cybersecurity. It's urgent to critically revise the foundations of machine learning focusing on open research questions that arise from practical requirements of cybersecurity applications and require novel, fundamental, understanding of machine learning theory;
2. Security of AI-based Tools and AI-empowered Systems: AI might become the weakest link of the cybersecurity chain, with a large increase of the attack surface. New algorithmic solutions

and practical software tools need to be developed for the secure design of AI-based tools and cyber-physical systems made up of AI-based and non-AI-based components;

3. MLSecDevOps: current industrial platforms for developing ML algorithms lack functionalities and modules for evaluating the algorithmic security of machine learning. The SOS AI project will develop a prototype, open-source platform for ML Secure DevOps, for the secure development of software tools containing ML algorithms;
4. AI Cyber Range: there is an alarming scarcity of work that considers realistic environments where ML tools are placed inside a real IT infrastructure. As an example, no previous work considered a real Cyber Range for the security evaluation of AI-based malware detectors. The SOS AI project will develop novel tools for AI security evaluation in next-generation cyber ranges;
5. AI Cyber Challenge: it is urgent to improve the level of university education and the training of industrial workforce on AI Security. The SOS AI project will develop a training program on AI security as part of the SERICS National Cybersecurity Academy.

WP Breakdown Structure

The conceptual organization of the SOS AI project is depicted in the next Figure that also illustrates the main work-packages, their relationships, and the project path from use-inspired basic research to engineering and education.



WP1 - Foundations of AI Security

WP1 Description: Machine-learning algorithms are widely used for cybersecurity applications. In these applications, the learning algorithm must face intelligent and adaptive attackers who can carefully manipulate data to purposely subvert the learning process. As theoretical foundations

of machine learning have not been originally thought under such premises, learning algorithms have been shown to be vulnerable to well-crafted attacks, including test-time evasion and training-time poisoning attacks (also known as adversarial examples). WP1 is aimed to critically revise the body of knowledge and the state of the art of adversarial machine learning under the lens of the so-called Pasteur's quadrant (https://en.wikipedia.org/wiki/Pasteur's_quadrant) in order to focus on open research questions that arise from practical requirements of cybersecurity applications and require novel, fundamental, understanding of machine learning theory.

Task 1.1 - Foundations of AI security evaluation

Task 1.2 - Foundations of attacks against AI and defenses

Task 1.3 – Pasteur's quadrant: use-inspired basic research for AI security **(See the Section Open Calls)**

WP2 – Security of AI-based Tools and AI-empowered Systems

WP2 Description: this work-package leverages the results of WP1 and delivers novel algorithmic solutions and practical software tools for security evaluation and protection of AI-based tools and AI-empowered systems. WP2 will contribute to the advancement of the state-of-the-art in three ways: 1)the delivered algorithmic solutions and practical software tools will take explicitly into account the requirements of selected cybersecurity applications, overcoming the unrealistic assumptions of the majority of the solutions proposed so far (e.g., the practical feasibility of the attacks will be taken explicitly into account); 2)challenging and novel application domains will be considered, such as the cyber-physical security of computer vision for driver assistance systems; 3)we will go beyond the state of the art that considered the security of "isolated" machine learning algorithms by analyzing the security of larger, AI-empowered, cyber-physical systems made up of AI-based and non-AI-based components (e.g., malware detection architectures using black listing, machine-learning static analysis, etc.). The activities of this WP2 will be coordinated with the WP4 of the project SANDSTORM of SERICS Spoke 7.

Task 2.1 - Methods for security evaluation of AI-based tools and AI-empowered systems

Task 2.2 – Defenses of AI-based tools and AI-empowered systems

Task 2.3 – Use cases and experiments **(See the Section Open Calls)**

WP3 – MLSecOps and AI Cyber Range

WP3 Description: in most of the cases, state-of-art attacks against ML and the related defenses have been designed in laboratory conditions. For instance, adversarial attacks are often executed in the digital domain, paying little or no attention to their practical feasibility in the physical world. This means that the threat of such attacks against AI is often over or underestimated. On the other hand, while basic research is doing a lot of seminal work on AI security, companies are working on automating the development and operations of ML models (MLOps) without focusing too much on ML security-related issues. Current industrial DevOps platforms lack functionalities and modules for evaluating the algorithmic security of machine learning. This work-package aims to bridge these gaps by extending the current MLOps paradigm to also encompass ML security (MLSecOps) and adding functionalities for the security evaluation of AI-based malware detectors in next-generation cyber ranges.

Task 3.1 - Development of the SERICS MLSecOps platform

Task 3.2 – AI Security Evaluation in Next-Generation Cyber Ranges

Task 3.3 – Use cases, best practices and SERICS Guide for Security of AI **(See the Section Open Calls)**

WP4 – AI Cyber Challenge

WP4 Description: this work-package aims to develop a training program on AI security as part of the SERICS National Cybersecurity Academy.

Task 4.1 - Development of the AI Cyber Challenge platform

Task 4.2 – AI Cyber Challenge

Open Calls / Bando a Cascata

Project Scope S.O.S. AI

WP1 Task 1.3 – Pasteur's quadrant: use-inspired basic research for AI security

Task 1.3 Description and expected outputs: this task is aimed to go beyond the achievements of tasks 1.1 and 1.2 and addresses additional use cases of cybersecurity where the current state of the art of machine learning is not enough to properly consider the applications' requirements. To this end, the consortium looks for external expertise by Open Call. The applicants to the Open Call should propose additional use cases (e.g., image forensics, cyber-physical security, etc.) not already addressed in tasks 1.1. and 1.2 by the partners' consortium and deliver new theoretical insights and practical solutions. The output of task 1.3 should be a technical report and software packages that can be made public available in open-source libraries (e.g., the SecML library, https://secml.readthedocs.io). The applicants to the Open Call should propose relevant KPIs (key performance indicators) that they promise to deliver (number of papers published in relevant conferences and journals, etc.).

Task 2.3 – Use cases and experiments

Task 2.3 Description and expected outputs: this task is aimed to go beyond the achievements of tasks 2.1 and 2.2 by addressing relevant use cases of cybersecurity (e.g., image forensics, security of cyber-physical systems, etc.) and/or addressing open issues not already tackled in tasks 2.1 and 2.2 by the partners' consortium (e.g. how to exploit domain knowledge in data-driven machine learning to increase the robustness against deliberate attacks, how to evaluate the security of large cyber-physical/software systems containing a few machine-learning components to empower the system). To this end, the consortium looks for external expertise by Open Call. The outputs of task 2.3 should be a technical report and software packages that can be made public available in open-source libraries (e.g., the SecML library, https://secml.readthedocs.io). The applicants to the Open Call should propose relevant KPIs (key performance indicators) that they promise to deliver (number of papers published in relevant conferences and journals, etc.).

Task 3.3 – Use cases, best practices and SERICS Guide for Security of AI

Task 3.3 Description and expected outputs: this task is aimed to leverage and expand the results of WP2 and WP3 to summarize and revise critically the best practices learnt by the SOS AI project and compile the SERICS Guide for the Security of AI. The output of task 3.3 should be a technical report that constitutes the SERICS Guide for Security of AI. The applicants to the Open Call should propose relevant KPIs (key performance indicators) that they promise to deliver for promoting and disseminating the SERICS Guide for Security of AI.