AI for Cybersecurity: from Adversarial Anomaly Detection to Intelligent Network Security Systems

Francesco Bergadano, Idilio Drago
Dipartimento di Informatica, Università di Torino
francesco.bergadano@unito.it, idilio.drago@unito.it

Abstract

The speed in which attacks appear and propagate requires efficient cyber protections. New attacks are mounted with the support of modern Machine Learning (ML) and AI algorithms, e.g., by learning users' behavior from data. Fighting such attacks requires systems that can learn attack patterns efficiently and autonomously. ML and AI represent an opportunity to evolve cyber-defenses, too. Here we address and summarize our research efforts in this direction: (1) anomaly detection in cybersecurity using ML tools and in presence of adversaries, (2) classification and generation of strings that are relevant for cybersecurity, in particular passwords, subdomains and phishing URLs, and (3) AI-powered network security, including new representations for darknet traffic and adaptive honeypots.

1 Introduction

In this short paper we summarize the research activities carried out at the Department of Computer Science of the University of Torino, concerning the application of ML and AI to several cybersecurity problems. In particular, we describe how we are using these algorithms for anomaly detection in presence of adversaries, for the generation of synthetic data that can support cybersecurity tasks (e.g., feeding AI models to fight attacks) and the use of AI for running systems supporting network security.

2 Adversarial Anomaly Detection

At the Department of Computer Science of the University of Torino we have done pioneering work on the applications of ML to ICT security, especially in the areas of anomaly detection and user authentication [Bergadano *et al.*, 2002; Bergadano *et al.*, 2003]. In particular, we have proved that keystroke dynamics can be used effectively to recognize legitimate users. This can be used for initial user authentication, for continuous and multi-level authentication, and for anomaly detection. The data we have used include keystroke duration and timing between one keystroke and the next. Every time a user strikes or releases a key, a time measurement is obtained, as well as the identifier of the key. A trigraph, composed of three keyboard identifiers, is then associated to six time measurements, that are then turned into corresponding delays or time-differences. We have shown that such trigraph-based delays can help learn highly accurate user classifiers.

More recently, we have been studying and applying Machine Learning to cybersecurity in an adversarial context [Bergadano, 2020].

In particular, we have addressed the problem of $evasion$, i.e., the possibility for an adversary to replicate the learning process, or figure out its outcomes, and evade defenses based on that knowledge. For example, in an intrusion detection scenario, the adversary could devise attacks that escape detection. The adversary could for example generate a large number of attack variants, and select the ones that are not recognized by the classifier.

A common contermeasure to evasion is randomization [Biggio *et al.*, 2008; Rota Bulò *et al.*, 2017]: the classifier is made unpredictable by introducing random components in a number of possible ways. In [Bergadano, 2019], we have developed a general framework for classifier randomization, that we have named "keyed learning", based on the idea that a secret "key" is used in different moments of the learning process. The adversary, not knowing the key, is unable to replicate the learning process, and cannot evade detection based on classifier knowledge.

Our keyed learning framework was applied to a practical problem that has recently received increased attention in cybersecurity: the detection of Web Application defacement. The method proved effective on a large-scale experimentation based on data provided by a major Italian manufacturing company [Bergadano *et al.*, 2019].

We now are addressing the next challenge that we view as important: finding out whether the keyed learning approach to randomization can provide superior results, if compared to simpler methods based on the introduction of random noise into the learned classifier. This question, for which we do not have an answer at the moment, can be approached both theoretically, using simple classifiers (e.g., linear or polynomial discriminants, or simple decisione trees), and experimentally, by testing the two approaches on well-know data sets, and possibly on new data sets that are particularly adequate to this purpose.

3 Data Generation for Cybersecurity

Automatic data generation can support several cybersecurity applications. Examples include the generation of strings to automatically test applications, as well as the creation of artificial datasets that augment AI models during training.

Here we illustrate the concept with three cases: automatic password checking, domain generation and phishing URL generation.

3.1 Password Checking

A password can be viewed as a string, and deciding whether a password is strong or weak can be regarded as a string classification problem. Typically, a password is considered to be weak if it is a simple mutation of a string belonging to a given dictionary [Bergadano *et al.*, 1998].

One could simply do a dictionary check, but this may be time and space consuming. In [Bergadano *et al.*, 1998], we have transformed a dictionary into a decision tree classifier, using standard ML algorithms. The resulting classifier could then tell whether a password is weak with reasonable approximation and with high efficiency and dictionary compression.

More recently, we have shifted our attention to the use of Generative Adversarial Networks (GANs) for similar applications. It has been shown that a GAN can generate a large number of passwords, starting form a dictionary (see, e.g., PASSGAN [Hitaj *et al.*, 2017]). PassGAN, using the publicly available dictionary "rockyou", was able to recognize 24,2% for the set of passwords obtained from the Linkedin password leak, though the two data sets are not directly related.

3.2 Subdomain Enumeration

Subdomain enumeration is used both in defensive and offensive approaches to cybersecurity. As an offensive tool, it can help discover hosts, naming policies, and vulnerabilities in the target perimeter. Subdomain enumeration may also be the basis for subdomain takeover. As a defensive control, it can help devise subdomain names that are difficult to detect, or list subdomains that should be given additional protection because of their visibility.

Enumeration is followed by validation (check for actual existence), but if this is done exhaustively it can be time consuming and hard to complete. Instead, we use GANs to generate subdomain candidates, using an innovative methodology and providing an extensive experimentation that demonstrates the practical usability of the method [Degani *et al.*, 2022]. This research is being carried out in collaboration with the University of Trento.

3.3 Phishing-Squatting Domain Generation

Phishing is a cyber attack in which attackers steal personal information from users using fraudulent messages. Email, SMS, and instant message apps have all been used as vectors to carry phishing attacks. Phishing messages often include URLs to bring victims to attackers' systems where the fraud actually takes place. The latter can be, for instance, a clone of login pages of famous services or fraudulent payment systems. Phishing is usually coupled with cybersquatting, in which attackers register domain names that are similar to legitimate services to increase the probability of fooling victims. The protection against phishing is achieved by filtering messages and by blocking malicious URLs. Both approaches however suffer from limitations.

The identification of phishing content is effective when phishing is distributed in massive campaigns (i.e., as spam). Spear phishing, i.e., well-crafted phishing targeting specific groups, is however much harder to be identified. Similarly, blocking cybersquatting URLs requires the creation of block lists that include new phishing/squatting domains as they are registered. The management of such block lists is time-consuming and often manual and, as such, inefficient against zero-day attacks.

Using approaches similar to the subdomain enumeration problem, we are developing new approaches to fight phishing and cybersquatting using AI [Valentim *et al.*, 2021; Trevisan e Drago, 2019]. The key idea is to use data augmentation to *generate* new samples of phishing content and cybersquatting URLs proactively, thus challenging (and automatically improving) phishing detection system. In particular, we are focusing on the use of GANs and Transformers, with a case study on the generation of phonetic phishing URLs. Our preliminary results, obtained in collaboration with UNIBS and Polito, show that AI algorithms can generate realistic phishing URLs with high potential to assist in the update of block lists.

4 AI-Powered Network Security

Understanding network attacks is a key step for network security. The attacking surface has grown throughout the years as more and more services are offered online. Automating security activities is paramount as a consequence, both to scale up the protection systems that must handle large data volumes as well as to speed up the reaction against zero-day attacks.

We are researching multiple ways to automate and scale up network security tasks using ML and AI [Trevisan *et al.*, 2020b; Trevisan *et al.*, 2017; Trevisan *et al.*, 2020a]. In the following, we provide some details of two cases that demonstrate the benefits of the approach.

4.1 New Representations for Darknet Traffic

Darknets (or network telescopes) are sets of IP addresses advertised by routing protocols without hosting services. Darknets have been used as passive sensors for monitoring attacks in the Internet, including the spread of botnets and the activity of scanners. Darknets receive a huge amount of packets, including benign or uninteresting traffic, such as traffic from scanners searching for well-known services and traffic due to misconfigurations.

We have shown that darknet traffic arrives from hundreds of thousands of sources, and various groups of sources collaborate to perform specific activities [Soro *et al.*, 2019; Soro *et al.*, 2020]. For example, multiple bots belonging to the same botnet usually collaborate to scan for services aiming to launch to a particular attack. The darknet observes the activity of these coordinated sources, however without clear indications on which traffic is due to particular groups or attacks. Identifying *relevant* events in darknet traffic is a complex task, which can however bring important insights to security analysts. We are working on methods to automate the analysis of darknet traffic with the support of AI. We have recently proposed Darkvec [Gioacchini *et al.*,], in conjunction with colleagues from Polito and Huawei. Darkvec is a system to automatically identify coordinated sources in darknet traffic. Darkvec includes a novel representation for darknet traffic, automatically learned from the packets. It is based on text mining approaches, in particular Word2Vec. It automatically unveils temporal and spatial correlations on traffic sent by different sources. We have demonstrated Darkvec in a case study using darknet traces, in which the IP addresses of multiple actors performing scans have been automatically uncovered.

Darkvec helps security experts and network operators to identify zero-day attacks, providing the means for protecting networks in a timely fashion.

4.2 Adaptive Honeypots

Honeypots have been used as a complementary source of data to obtain information about network attacks. Unlike darknets, honeypots answer unsolicited traffic, trying to engage with attackers to obtain details about their intentions and tools.

Honeypots are usually classified as high-interaction – when real systems are deployed to be attacked, or low-interaction – when a simulated environment is used instead. Both approaches have drawbacks and advantages. The former are realistic, but more risky and harder to operate. The latter are easier to be controlled and monitored, but suffer from higher development and maintenance costs and the lack of realism. Indeed, attackers often profile systems after the exploitation of a vulnerability, identifying and avoiding honeypots.

We are building novel adaptive honeypots, including (i) DPIPot, a honeypot that performs deep packet inspection to identify the best approach to use when facing unsolicited traffic [Rescio *et al.*, 2021], and (ii) CannyPot, a honeypot powered with Reinforcement Learning (RL) [Milan *et al.*, 2021] that learns strategies to react to attackers in autonomy.

Our CannyPot prototype, developed with Polito and Huawei colleagues, can speak SSH and is already deployed online for some months. The system observes terminal commands sent by attackers and learns how to react using multiple real backend systems running in a controlled environment. The system's RL engine guides the selection of strategies to respond to attacks, aiming to maximize the time attackers remain connected to the honeypot. Preliminary results demonstrate the advantage of CannyPot (currently under review).

5 Concluding Remarks

In this short article we have described multiple examples of results supporting the use of AI in cybersecurity applications. While a complete automation of cybersecurity tasks is still very unlikely in the short-term, given the complexity of the problem, we believe that the application of AI for cybersecurity will continue helping to reduce attacking opportunities, thus increasing the protection of systems.

Riferimenti bibliografici

[Bergadano *et al.*, 1998] Francesco Bergadano, Bruno Crispo, e Giancarlo Ruffo. High dictionary compression for proactive password checking. *ACM Transactions on Information and System Security (TISSEC)*, 1(1):3–25, 1998.

[Bergadano *et al.*, 2002] Francesco Bergadano, Daniele Gunetti, e Claudia Picardi. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):367–397, 2002.

[Bergadano *et al.*, 2003] Francesco Bergadano, Daniele Gunetti, e Claudia Picardi. Identity verification through dynamic keystroke analysis. *Intelligent Data Analysis*, 7(5):469–496, 2003.

[Bergadano *et al.*, 2019] Francesco Bergadano, Fabio Carretto, Fabio Cogno, e Dario Ragno. Defacement detection with passive adversaries. *Algorithms*, 12(8):150, 2019.

[Bergadano, 2019] Francesco Bergadano. Keyed learning: An adversarial learning framework—formalization, challenges, and anomaly detection applications. *ETRI Journal*, 41(5):608–618, 2019.

[Bergadano, 2020] Francesco Bergadano. Adversarial learning e le sue applicazioni nella cyber security. *ICT Security Magazine*, March 2020.

[Biggio *et al.*, 2008] B. Biggio, G. Fumera, e F. Roli. Adversarial pattern classification using multiple classifiers and randomization. In *Proc. Int. W. on Structural, Syntactic and Statistical Pattern Recognition*, 2008.

[Degani *et al.*, 2022] Luca Degani, Francesco Bergadano, Seyed Ali Mirheidari, Bruno Crispo, e Fabio Martinelli. Generative adversarial networks for subdomain enumeration. *ACM SAC Conference*, 2022.

[Gioacchini *et al.*,] Luca Gioacchini, Luca Vassio, Marco Mellia, Idilio Drago, Zied Ben Houidi, e Dario Rossi. Darkvec: Automatic analysis of darknet traffic with word embeddings. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '21, page 76–89, New York, NY, USA.

[Hitaj *et al.*, 2017] B. Hitaj, P. P. Gasti, G. Ateniese, e F. Perez-Cruz. Passgan: A deep learning approach for password guessing. *arXiv:1709.00440*, 2017.

[Milan *et al.*, 2021] Giulia Milan, Luca Vassio, Idilio Drago, e Marco Mellia. Rl-iot: Reinforcement learning to interact with iot devices. In *IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, pages 1–6, 2021.

[Rescio *et al.*, 2021] Tommaso Rescio, Thomas Favale, Francesca Soro, Marco Mellia, e Idilio Drago. Dpi solutions in practice: Benchmark and comparison. In *IEEE Security and Privacy Workshops (SPW)*, pages 37–42, 2021.

[Rota Bulò *et al.*, 2017] S. Rota Bulò, B. Biggio, I. Pillai, M. Pelillo, e F. Roli. Randomized prediction games for adversarial machine learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2466–2478, 2017.

[Soro *et al.*, 2019] Francesca Soro, Idilio Drago, Martino Trevisan, Marco Mellia, J. Ceron, e J. Santanna. Are darknets all the same? on darknet visibility for security monitoring. In *IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, pages 1–6, 2019.

[Soro *et al.*, 2020] Francesca Soro, Mauro Allegretta, Marco Mellia, Idilio Drago, e Leandro M. Bertholdo. Sensing the noise: Uncovering communities in darknet traffic. In *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*, pages 1–8, 2020.

[Trevisan *et al.*, 2017] Martino Trevisan, Idilio Drago, Marco Mellia, e Maurizio M. Munafo. Automatic detection of dns manipulations. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4010–4015, 2017.

[Trevisan *et al.*, 2020a] Martino Trevisan, Danilo Giordano, Idilio Drago, Maurizio Matteo Munafo, e Marco Mellia. Five years at the edge: Watching internet from the isp network. *IEEE/ACM Transactions on Networking*, 28(2):561–574, 2020.

[Trevisan *et al.*, 2020b] Martino Trevisan, Francesca Soro, Marco Mellia, Idilio Drago, e Ricardo Morla. Does domain name encryption increase users' privacy? *SIGCOMM Comput. Commun. Rev.*, 50(3):16–22, jul 2020.

[Trevisan e Drago, 2019] Martino Trevisan e Idilio Drago. Robust url classification with generative adversarial networks. *SIGMETRICS Perform. Eval. Rev.*, 46(3):143–146, jan 2019.

[Valentim *et al.*, 2021] Rodolfo Valentim, Idilio Drago, Martino Trevisan, Federico Cerutti, e Marco Mellia. Augmenting phishing squatting detection with gans. In *Proceedings of the CoNEXT Student Workshop*, CoNEXT-SW '21, page 3–4, New York, NY, USA, 2021.