Attack scenario generation using machine learning for IDS evaluation

- Advisors:
  - Ludovic Mé (ludovic.me@inria.fr)
  - Frédéric Majorczyk (frederic.majorczyk@irisa.fr)
  - Mouad Lemoudden (mouad.lemoudden@inria.fr)
  - Samuel Hangouet (samuel.hangouet@intradef.gouv.fr)
- Team: CIDRE (`https://www.inria.fr/en/teams/cidre`)

- keywords: intrusion detection, machine learning, attack scenario generation

Subject

Intrusion detection systems (IDS) are indispensable tools for the security of information systems. In general, it is impossible to completely secure a given information system: new vulnerabilities are discovered in software, protection measures may be too restrictive for the user, etc. These intrusion detection systems make it possible to detect violations of security policy and to report them to a security operator so that necessary measures can be carried out.

From a scientific and operational point of view, it is necessary to evaluate detection solutions: to understand their limits and to determine how to improve them. The evaluation of intrusion detection systems is a scientific issue in itself. This internship focuses on a sub-issue of the global problem, which is the generation of an attack scenario.

An intrusion detection system can be considered as a binary classifier: an action on the system can either be malicious or not, and subsequently, the IDS responds by raising an alert or not. Conventional measures such as the false-positive rate (i.e. false alerts) and the false-negative rate (i.e. absence of alerts in case of malicious action) are indicators very often measured in the state of the art.

To measure the false negative rate, it is necessary to generate malicious activities on the monitored system. Moreover, to obtain representative measures, the malicious activities should have characteristics from real malicious activities in the wild; they also should have goals that are realistic with relation to the monitored system and should cover a maximum of unitary attacks.

Classical state-of-the-art attack scenario generation techniques [6, 4] are based on the complete knowledge of the information system (topology, configuration, vulnerabilities, etc.) and on attack graph generation (with nodes representing a state of the system). Their objective was not to generate realistic malicious activities for IDS evaluation but to analyze the security of the information system considered and propose solutions to protect the system of the discovered attacks.

Recent advances in machine learning techniques allow to consider their use to generate attack scenarios that try to mimick the attacks in the wild. For example, the model of reinforcement learning [7, 5] is based on an agent that observes his environment, chooses and does an action, and gets a reward (that can be negative). This generic model seems to match this case of the internship: the agent can model an attacker that tries some actions on the system and gets rewards when they are successful. It would certainly be necessary to modify the model to take into account the fact that the attacks should be representative of attacks happening in the wild.

The MITRE Att&ck Framework [1] can be a source for the different attack steps that the agent can take on the system. This framework models attackers' behavior and the platforms that attackers' group are known to target. Some tools, such as Caldera [2] and Metta [3], use this framework to do adversarial simulation so as to, for example, train blue teams (i.e. teams that defend information system).

During this internship, we want to study the different types of machine learning techniques able to generate attack scenarios, to define the inputs of these algorithms, to carry out learning on a case study, to generate some attack scenarios for our case study.

References

[1] Att&ck. `https://attack.mitre.org/wiki/Main_Page`.

[2] Caldera. `https://github.com/mitre/caldera`.

[3] Metta. `https://github.com/uber-common/metta`.

[4] K. Ingols, R. Lippmann, and K. Piwowarski. Practical attack graph generation for network defense. In *Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual*, pages 121–130. IEEE, 2006.

[5] S. Kelly and M. I. Heywood. Emergent tangled graph representations for atari game playing agents. In *European Conference on Genetic Programming*, pages 64–79. Springer, 2017.

[6] X. Ou, W. F. Boyer, and M. A. McQueen. A scalable approach to attack graph generation. In *Proceedings of the 13th ACM conference on Computer and communications security*, pages 336–345. ACM, 2006.

[7] R. S. Sutton and A. G. Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.