Title: Adversarial Vulnerability Measurement of AI-boosted Intrusion Detection Systems

Subjects of the project:
The task of Intrusion Detection Systems (IDS) is to flag potentially malicious system usages or network traffic by analyzing system event logs and network traffic flows. The detection output of IDS services helps monitor the health status of key IT infrastructures and mitigate potential threats. AI-boosted automated IDS methods, such as AI-based anomaly detection approaches, have become more and more popular in both research communities and industrial practices. Unlike the classical signature-based IDS systems, AI-boosted approaches can automatically capture the suspicious behavioral profiles drifted significantly from normal usage and report potential attacks with a high precision.

Effective as it is, the AI technology is also well known to be prone to adversarial attacks. By slightly perturbing the input to AI models, an adversary can drastically change their decision outputs. This raises a severe trustworthy issue over the AI-based IDS applications. For example, an adversary can modify the attack behaviors to evade the detection of AI, which causes high false negative rate scores and make AI detection practically infeasible.

In this project, we will first review the mainstream adversarial attack methods against AI systems, such as Fast Gradient Sign Method (FGSM) [1], Projected Gradient Descent (PGD) [2] and multi-armed bandit [3]. These methods can be applied to both numerical and categorical data. We will use these three methods to conduct adversarial attacks against AI-based anomaly detection methods used in IDS applications. In our work, we will organize a systematic measurement study to test the adversarial vulnerability of 3 state-of-the-art AI-based anomaly detection methods, namely DeepLog [4], Tiresias [5] and DeepCase [6] over 2 publicly available large-scale system log datasets, HDFS [4,6] and IPS-Ping [5]. Specifically, we will compare the accuracy of each of the anomaly detection methods before and after being attacked. We will specially focus on False Alarm Rate (FAR) and False Negative Rate (FNR) to evaluate the detection robustness of the AI-based anomaly detection methods.

In this project, we will:

1. **review the mainstream adversarial attack methods against AI systems**, such as Fast Gradient Sign Method (FGSM) [1], Projected Gradient Descent (PGD) [2] and multi-armed bandit [3]. These methods can be applied to both numerical and categorical data.
2. **conduct adversarial attacks against AI-based anomaly detection methods** used in IDS applications. In our work, we will organize a systematic measurement study to test the adversarial vulnerability of 3 state-of-the-art AI-based anomaly detection methods, namely DeepLog [4], Tiresias [5] and DeepCase [6] over 2 publicly available large-scale system log datasets, HDFS [4,6] and IPS-Ping [5].
3. **compare the accuracy of each of the anomaly detection methods** before and after being attacked. We will specially focus on False Alarm Rate (FAR) and False Negative Rate (FNR) to evaluate the detection robustness of the AI-based anomaly detection methods.

Contact Information

The research intern will be hosted and conduct the projects at INRIA Rennes, campus universitaire de Beaulieu, Rennes, France. This project will be co-supervised by Cedric Gouy-Pailler (cedric.gouy-pailler@cea.fr) and Yufei Han (yufei.han@inria.fr). Should you have any further questions regarding the project, please don't hesitate to contact us via the two email addresses.

References

[1] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, Explaining and Harnessing Adversarial Examples, ICLR 2015.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018.

[3] Auer, P, "Using upper confidence bounds for online learning". Proceedings 41st Annual Symposium on Foundations of Computer Science. IEEE Comput. Soc. pp. 270–279

[4] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17). Association for Computing Machinery, New York, NY, USA, 1285–1298.

[5] Yun Shen, Enrico Mariconti, Pierre Antoine Vervier, and Gianluca Stringhini. 2018. Tiresias: Predicting Security Events Through Deep Learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18). Association for Computing Machinery, New York, NY, USA, 592–605.

[6] Thijs van Ede, Hojjat Aghakhani, Noah Spahn, Riccardo Bortolameotti, Marco Cova, Andrea Continella, Maarten van Steen, Andreas Peter, Christopher Kruegel, Giovanni Vigna, DEEPCASE: Semi-Supervised Contextual Analysis of Security Events,