Towards an Information-Theoretic Framework of Intrusion Detection for Composed Systems and Robustness Analyses - *Presentation abstract*

Tobias Mages[a*], Christian Rohner[a] and Magnus Almgren[b]

[a]*Department of Information Technology, Uppsala University, 752 36 Uppsala, Sweden*

[b]*Department of Computer Science and Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden*

Abstract

Network-based Intrusion Detection Systems (NIDSs) face among others the combined challenge of maintaining a high Positive Predictive Value (PPV) at low base-rates (Axelsson, 1999) and being resilient against adaptive adversaries (Hashemi et al., 2019; Hartl et al., 2020; Sheatsley et al., 2020). Due to these challenges, modern IDSs are often composed of several detection methods or systems. This can enable a reduction of false alarms at low base-rates (Meng and Kwok, 2013), diversify the used feature levels to increase robustness or enable utilizing the advantages of different detection methods. However, currently there exists no framework that would be suitable for a detailed analysis of such composed systems in the area.

In this work, we present an Information-Theoretic Framework (ITF) for the evaluation of composed detection systems. It extends the ITF from Gu et al. (2006) to enable a detailed analysis and comparison of IDSs by quantifying the performance of its individual components, including multiple feature representations, detection methods and their specific arrangement. This also enables studying the dynamics between operation points to fine-tune parameters or evaluate threat models and attack methods by analyzing the robustness dependencies between different classifiers within the system.

We demonstrate the versatility of the framework by showing the impact of an evasion attempt with adversarial examples on a composed IDS. In particular we attribute the overall system performance to its individual components, show the impact of compositions on their operation points and make statements about the system performance at different base rates. The results motivate how even a small fraction of benign samples under adversarial control could render an IDS ineffective and indicate that existing classification redundancies might not be fully utilized during an attack due to a static system design.

References

S. Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, CCS '99, page 1–7, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131488. doi: 10.1145/319709.319710. URL `https://doi.org/10.1145/319709.319710`.

G. Gu, P. Fogla, D. Dagon, W. Lee, and B. Skoric. Towards an information-theoretic framework for analyzing intrusion detection systems. In D. Gollmann, J. Meier, and A. Sabelfeld, editors, *Computer Security – ESORICS 2006*, pages 527–546, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-44605-7.

[*]*Presenter*

A. Hartl, M. Bachl, J. Fabini, and T. Zseby. Explainability and adversarial robustness for rnns. In *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 148–156, 2020. doi: 10.1109/BigDataService49289.2020.00030.

M. J. Hashemi, G. Cusack, and E. Keller. Towards evaluation of nidss in adversarial setting. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, Big-DAMA '19, page 14–21, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369992. doi: 10.1145/3359992.3366642. URL https://doi.org/10.1145/3359992.3366642.

Y. Meng and L. Kwok. Towards an information-theoretic approach for measuring intelligent false alarm reduction in intrusion detection. In *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pages 241–248, 2013. doi: 10.1109/TrustCom.2013.33.

R. Sheatsley, N. Papernot, M. Weisman, G. Verma, and P. McDaniel. Adversarial examples in constrained domains. *arXiv preprint arXiv:2011.01183*, 2020.