Poster: Sometimes, You Aren't What You Do: Mimicry Attacks against Provenance Graph Host Intrusion Detection Systems

Akul Goyal
University of Illinois at Urbana-Champaign
akulg2@illinois.edu

Xueyuan Han
Wake Forest University
vanbasm@wfu.edu

Gang Wang
University of Illinois at Urbana-Champaign
gangw@illinois.edu

Adam Bates
University of Illinois at Urbana-Champaign
batesa@illinois.edu

***Abstract*—Modern-day provenance-based threat-hunting tools have become a valuable asset in the repertoire of system analysts. Given a system alert, threat-hunting tools guide the system analyst during attack forensics to recover the entire attack footprint of the adversary. However, many of these tools face an onslaught of false positive alerts from static rule sets used in enterprise settings. Many threat-detecting systems combat this by ranking outputted alerts so system analysts focus on alerts correlating to attack activity. In [1], Goyal et al. showed that the current landscape of Prov-IDSs is susceptible to evasion against an active attacker who introduces additional benign behavior alongside their attack. We look to extend this work by applying a similar technique to many modern-day threat-hunting tools. Our goal is to evaluate the efficacy of these systems when paired up against an adaptive attacker that executes innocuous activity alongside their malicious behavior. Our results show that one of the quintessential threat-hunting tools is vulnerable to attack, mis-ranking alerts such that 100% of the false positive alerts occur before the true positive alerts.**

***Index Terms*—intrusion detection system, endpoint detection and response, threat hunting**

I. Introduction

Automatically detecting malware has always been a challenging problem in computer security - dating back to the start of 1972 with the release of the Computer Security Technology and Planning Study. Anomaly-based intrusion detection systems (IDS) have become an avenue of malware detection, training on benign data to model normal system behavior. Host intrusion detection systems (HIDS), particularly provenance-based HIDS (Prov-IDS), are a subset of IDS that utilize audit logs to detect abnormal behavior within the host system. Provenance graphs represent audit logs in a graphical structure where every node represents a system entity, and edges correspond to the system calls the host operating system makes between the two system entities. Provenance graphs provide many benefits to intrusion detection, particularly preserving the locality between related system interactions regardless of their temporal distance. As a result, Prov-IDS harbor a "natural" resistance to Advance Persistent Attacks (APT) where the attacker conducts each step of their attack step over long periods on the system.

Provenance graphs also aid many threat investigation tools required to reconstruct the attacker's footprints given an alert. These systems utilize the causal relationship between system events to backtrack and uncover the exact steps conducted by the attacker to gain access to the victim system. Many threat investigation tools also combat alert fatigue by utilizing graphical properties (edge frequencies) that prioritize "true" alerts to the system analyst.

Goyal et al. citegoyalsometimes investigated how current Prov-IDS are susceptible to evasion against an active attacker that inserted additional benign behavior to the provenance graph alongside their attack. This work showed that because Prov-IDS utilized off-the-shelf machine learning techniques to summarize graphs, they lost critical relationships within the graph that allowed the attacker to evade detection successfully. A similar but orthogonal property exists within modern-day attack reconstruction and threat-hunting systems that utilize provenance graphs. These systems make strong assumptions about the attacker's behavior, making them brittle when faced with an active attacker. In this poster, we extend the mimicry attack gadgets introduced in [1] to the landscape of endpoint detection and response (EDR) and threat-hunting tools.

II. Related Work

Limited work has occurred in investigating the efficacy of modern-day threat-hunting tools. Goyal et al. [1] evaluated the robustness of Prov-IDS against an active attacker but did not consider threat-hunting tools. Several previous works [2] from the adversarial machine learning community have also successfully evaded graph-based learning systems. However, many of these works make strong assumptions about the systems they attack (neural networks) that prevent them from being applied directly to current threat-hunting tools. Goyal et al. also show that black box graph-based adversarial attacks are inefficient in creating the changes required to evade detection against threat-hunting tools. This work is the first to consider attacks against threat-hunting systems.

III. Methodology and Results

A. Background

This work will investigate the efficacy of a canonical threat investigation system: NoDoze [3]. For the brevity of this poster abstract, we will briefly describe the system - focusing on the parts open to adversarial perturbation. We encourage the reader to look at the paper for a better understanding. Hassan et al. designed NoDoze to reflect system analysts' behavior of defining the suspiciousness of a specific alert by its historical context. At train time, NoDoze generates an edge frequency table based on benign-only graphs representing normal activity on the host system. Next, for each alert provided by a given Prov-IDS, NoDoze creates a dependency graph. NoDoze generates this dependency graph by conducting a forward and backward trace in the provenance graph from the provided alert. Each edge $e = (v_i, r_j, v_k)$, representing a system relationship (system call) between two system entities, is assigned a transition probability $tr(e)$ according to the following equation $tr(e) = \frac{Freq(e)}{Freq_s(e)}$ where $Freq(e)$ is the frequency of the edge and $Freq_s(e)$ is the source frequency or how many times the source (v_i), relationship (r_j) pair occur in the training dataset. All possible paths within the dependency graph are then assigned an anomaly score representing the sum of each edge transition probability within the given path. NoDoze aggregates the most k most suspicious paths and assigns an overall anomaly score to the given alert—alerts with a higher score ranker to the system analyst for investigation.

B. Attacking Threat Investigation

NoDoze makes many strong assumptions within its system design that allow an attacker to evade detection. Assuming that the goal of an attacker is to prioritize false positive alerts such that the system analyst is less likely to uncover the attack, the attacker can include additional behaviors to reduce the overall rarity score of the given attack. We assume an attacker that cannot change their attack footprint (modify the attack path) and has control of a process on the host system, allowing them to execute additional behavior under the scope of the given process's privileges. We also assume that the attacker knows the distribution of anomaly scores within the false positive alerts. While this may seem like a strong assumption, the attacker can simulate the victim's activity on a surrogate system and generate a reasonable estimate for the false positive anomaly score distribution.

Like many Prov-IDS evaluated in [1], NoDoze must compare against paths of different sizes. NoDoze handles this by introducing a decay factor within the summation of a path's anomaly score. As a result, an activity occurring earlier in a given path carries a smaller weight on the overall path's anomaly score than an activity that happens later. An attacker aware that NoDoze is running on the system can frontload their attack and introduce benign activity afterward. As a result, the abnormal parts for the attack path have a decreasing weight on the overall summation of the path's anomaly score, while the benign (high-frequency edge) occurring later on by the attacker has a more substantial effect on the overall path's anomaly score.

C. Results

To validate our attack methodology, we experimented against the StreamSpot dataset [4]. This dataset contains 600 provenance graphs, with 500 of the graphs representing benign activity and 100 of the graphs representing attack activity. We split the dataset such that NoDoze is trained on 375 benign graphs, validated against 25 benign graphs, and tested against 100 attack and benign graphs. For each of the 100 attack graphs, we run a commercial EDR that fires alerts using static rules. We pick a single alert that correlates to each graph's attacker behavior. Similarly, we run the EDR on the 100 benign graphs within the test set and identify a single alert - ensuring we feed in a false positive to NoDoze. We sample edges from the frequency table generated during training to guide the attacker on additionals behavior to insert alongside the attack - favoring high-frequency edges to complete the path. We reduced each alert anomaly score for each of the 100 attack graphs such that NoDoze prioritized the 100 benign graphs before ranking any of the attack graphs.

IV. Conclusion

Threat-hunting tools have provided an essential resource to system analysts who face an ever-growing number of false alerts. More importantly, threat-hunting tools are more general than Prov-IDS, being able to integrate with any system that provides an alert. As a result, it is important to evaluate many modern-day threat-hunting tools as they can affect many real-world systems. The consequence of this work showed that one of the most canonical examples of a threat-hunting tool is vulnerable to an active adversary who introduces additional benign behavior alongside their attack. Future work can expand to other threat-hunting tools like Atlas [5] and RapSheet [6].

Acknowledgment

This work was supported by NSF under contracts CNS-16-57534, CNS-17-50024 and CNS-20-55127.

References

[1] A. Goyal, X. Han, G. Wang, and A. Bates, "Sometimes, you aren't what you do: Mimicry attacks against provenance graph host intrusion detection systems," in *network and distributed systems security symposium*, 2023.

[2] L. Sun, Y. Dou, C. Yang, J. Wang, Y. Liu, P. S. Yu, L. He, and B. Li, "Adversarial attack and defense on graph data: A survey," *arXiv preprint arXiv:1812.10528*, 2018.

[3] W. U. Hassan, S. Guo, D. Li, Z. Chen, K. Jee, Z. Li, and A. Bates, "Nodoze: Combatting threat alert fatigue with automated provenance triage," in *network and distributed systems security symposium*, 2019.

[4] E. Manzoor, S. M. Milajerdi, and L. Akoglu, "Fast memory-efficient anomaly detection in streaming heterogeneous graphs," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1035–1044.

[5] A. Alsaheel, Y. Nan, S. Ma, L. Yu, G. Walkup, Z. B. Celik, X. Zhang, and D. Xu, "Atlas: A sequence-based learning approach for attack investigation." in *USENIX Security Symposium*, 2021, pp. 3005–3022.

[6] W. U. Hassan, A. Bates, and D. Marino, "Tactical provenance analysis for endpoint detection and response systems," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 1172–1189.