Review #4079B

[Shuai Hao] R1 · May 12

Paper summary

The paper proposes an end-to-end system, OCR-APT, to detect APT attacks and reconstruct their investigation reports from system-level audit logs. The proposed framework leverages Graph Neural Networks (GNNs) and one-class SVM for anomaly detection, and designs carefully designed multi-stage prompts to reconstruct the attack stories through LLM. The results outperform previous detectors by achieving 0.95 precision and 1.00 recall on datasets such as DARPA TC3 and NODLINK.

Strengths / Reasons to accept

+ The paper proposes an end-to-end system to construct human-readable APT reports through provenance (sub)graphs and LLMs. Detailed prompt design and text are provided.

+ The paper presents a comprehensive evaluation, including both the framework's effectiveness and the validity of the LLM-generated report. The results demonstrate improvements compared to SOTA.

Weaknesses / Reasons to reject

– The paper lacks some detailed descriptions to understand certain technical aspects explicitly. See detailed comments below.

– The paper somewhat presents two separate components, the subgraph construction and LLM-empowered investigation. Given that the prior study has achieved comparable performance (e.g., FLASH, as shown in Table 2), feeding their results into LLM may produce similar results.

Constructive comments for author

– Figure 3 is not explicit on how the stages are converted. Specifically, it's hard to see how the nodes between stages 2 and 3 are mapped.

– Also, it's unclear how the subgraph partition is conducted and whether it impacts the detection results and attack investigation. Does the 4th subfigure in Figure 3 show partitioned subgraphs or not?

– In section 2.1, the paper stated that "provenance graphs are heterogeneous ... a malicious process node may still be correctly classified as a process, undermining anomaly detection." This causation is not clear. This kind of issue is the ineffectiveness of the classification model, rather than an inherent disadvantage of the methodology. A heterogeneous-based approach can still misclassify a malicious process as a benign one, undermining anomaly detection.

– In section 6.2, the paper stated that "the system validates the extracted IOCs to remove hallucinations". How does the system validate IOCs, and what are the criteria here to identify them?

– In section 5.1.1, the proposed OCRGCN avoids using IP addresses or file paths. However, it seems that it causes missed detection for OpTC51 in section 7.3, although it is identified by the following LLM stage (based on (IP addresses of) C&C server?). So does it indicate they are actually useful features?

– In addition, in the above case, it remains unclear how the LLM can offer such an enhanced detection performance. If the anomaly nodes are not detected, their associated subgraphs may not be provided to LLM.

– The evaluation only selects three hosts with the highest attack activities in OpTC case. Does this sufficiently evaluate and demonstrate the effectiveness of the proposed scheme? Probably evaluating the detection of malicious nodes with different activity levels is a more convincing scenario.

Questions for authors' response

– Please briefly clarify the aforementioned technical details, including the subgraph partition, IOCs validation, and feature selection.

– In the training phase, it looks like only two key features are used for behavior modeling, action frequencies and idle period. Can these two fundamental features sufficiently identify malicious behaviors for each type of node? Please briefly justify.

– The first step of subpragh construction is to identify direct connections between anomalous nodes. Does it mean the approach of subpragh construction only effective in the scenario with a single attack? If multiple malicious actions are present, the node-based approach may identify them individually but the subgraph may incorrectly connect different attacks, confusing the attack detection as well as the investigation report.

Does the paper raise ethical concerns?

1. No

Reviewer expertise in this domain

2. Some familiarity (I am aware of some work in this domain)

Reviewer confidence

2. Somewhat confident

Overall merit

4. Weak Accept