Linking CVE's to MITRE ATTCK Techniques - Notes

Cameron Noakes

November 2021

1 Sentence Overview

The researchers use the CVE regular expression dataset in order to obtain commonly exploited CVEs and vulnerabilities.

This paper used mathematics to obtain the data required to develop the model of a multi-head joint embedding neutral network design. Samples of 200 CVEs found in threat reports with their corresponding ATTCK techniques to be able to extract the context phrases.

2 Abstract - Notes

Clarification to the terminology of CVE, explaining that each CVE is tied to a software version and assigned number (version number) to allow exploitation of services and content management systems (CMS) software by threat actors.

2.1 Model Proposition - Notes

This paper proposes a model of Multi-Head Joint Embedding Neural Network to automatically map common vulnerabilities and exposures (CVE's) to the Mitre Attack techniques. This proposed model would allow easier mapping to specific attacks that were conducted to execute CVE's on known vulnerabilities and help mitigate known vulnerabilities since the Mitre Attack Framework has more explanations when it comes to mitigation strategies than places like Exploit-DB (a place where CVE's are stored).

It would also allow the grouping of CVE's by the techniques outlined in the Mitre Attack Framework to help quantitatively re-prioritize vulnerability risks according to attack stage defined in Mitre Attack Framework

3 Introduction - Notes

Statements of how the most commonly used CVEs are the ones that threat actors often use for exploitation, there is a select few hundred that attackers seem to always use. The researcher surveyed 690 advanced cyber threat reports from past 10 years and concluded that less than 500 CVE's are actively exploited.

This knowledge was said to be used to help defensive intrusion detection systems (IDS) to know what to commonly look for and where to target its resources.

The introduction continues to elaborate on how the most common exploited CVEs are ones relating to the top 100 technologies, client side software and web applications, allowing researchers knowledge of where to focus certain defensive security operations.

3.1 The Proposed Automation Model

The model will be mapped with reference to the dataset of CVE data collected from the past 10 years, allowing accurate, real time, up-to-date threats and attacks to be used to develop each component of the model, creating an overall more developed, industry-grade model to help map CVEs to the Mitre Framework Attack techniques.

the introduction continues with an outline of the sections for which each section discusses the relevant research and topic in turn.

3.2 Motivating Example

A specific example is used to discuss how mapping CVE's to Mitre techniques can be used as mitigation strategies to deploy intrusion detection systems to analyze network traffic for suspicious activity to help reduce the chances of an attackers successful threat.

4 Labelling Process

The researchers use the CVE regular expression dataset in order to obtain commonly exploited CVE's and vulnerabilities. However, this was outlined in the

'what problems does this solve' section of this document about how the researcher said that this dataset does not account for missed connections between the nature of techniques that are introduced by each CVE, since the model is based on this dataset it conclusively ascertains that the model will have gaps in accuracy much to the same effect as the dataset.

This paper uses mathematics to obtain the data required to develop the solution model, the researcher initially starts this process by manually labelling randomly sampled 200 CVE's found in threat reports with their corresponding Mitre techniques to be able to extract the context phrases from each mentioned CVE.

The paper then continues by discussing how the researcher measured L2 distance, Cosine distance, Maximum Mean Discrepancy (MMD) and Fisher Linear Discriminant (FLD) and then proceeded to select the best performing distance measurement method for labeling pipeline to continue to help map to Mitre techniques.

The final resolution for the mathematics is an equation based on the previously discussed factors to be able to find cosine similarities, the highest cosine similarity with the CVE phrases are assigned the label as a technique.

Mathematics was also used to develop the model, a multi-head deep embedding model for each CVE to predict corresponding ATTCK techniques.

5 Datasets

Numerical values were given in the paper about the attained raw data sets, 690 cyber security articles from a collection of advanced persistent threats (APT) reports, zero-day exploits observed, 63720 vulnerability reports and 37,000 threat reports identified for further data analysis to base the proposed solution on.

This section then detailed a popular security tool for vulnerability assessments called 'Nessus' which is an open source vulnerability scanner with many built-in features and plugins to help aid the security professional.

Since the vulnerability scanner can output information of related CVEs on a

scan, the result contains human-curated rich descriptions about the CVE, aiding in the mapping process as more data is returned to identify which technique the CVE should be mapped to.

6 Related Work

This section explained how new and old vulnerability information is mixed which can lead to biased results and due to concept drift it can make the solution models less robust.

"There has been some work into exploring the behavior of different attackers, but no work to the best of our knowledge has considered using attacker techniques and defender constraints in concert with machine learning for the purposes of vulnerability or software service management."

7 Research Gaps and Questions

A research gap for the proposed solution model of mapping CVE's to techniques in the Mitre Framework is that the regular expression (RegEx) dataset for CVEs is used to build the development of the model, the researcher outlined that this dataset could potentially miss the connection between the nature of the technique and its correlated CVE, making a less accurate solution since the model is based on an inaccurate dataset the final model will also have inaccurate representation of the data.

In the abstract they also call known vulnerabilities 'common vulnerabilities and exposures' (CVE's) which is not strictly accurate and are often said to be the exploit itself for such a vulnerability, not the vulnerability itself.

8 What problem does this solve?

CVEs over the past 10 years don't have a complete understanding for which category the CVE should be in, which can result in poor mitigation strategies and no direct mapping to detailed rich-text descriptions for attacks.

There is not many practices of relating specific CVEs to Mitre Attack Framework techniques, the closest solution is currently to link CVE's in threat reports to Mitre techniques using a series of regular expressions "which may miss the connection between the nature of techniques leveraged introduced by the CVE."

9 How do they solve it?

The proposed model will automatically map known CVE's to the Mitre Attack Framework and under a specific technique using a regular expression dataset and mathematical analysis of information in threat reports.

10 How do they assess it?

No mention of how they will assess it but in future work they mentioned how they would compare to bench marking of the original dataset.

11 Conclusion - Notes

Outlined the multi-head joint embedding neutral network designed in the paper and the datasets.

Any future work and some limitations were also discussed but weren't too relevant or have covered them in other sections.