Heterogeneous Network Mining of the National Vulnerability Database

Andrew Driggs
Introduction

• As more data, processes, and interactions are moved onto the Internet, the risks of vulnerable software leading to exposures increases

• These vulnerabilities and exposures are sometimes found by "the good guys" doing security research, and sometimes they are found by people looking at logs that notice intrusions

• Once these vulnerabilities and exposures are discovered, they are described in a report that gets added to the NVD for distribution

• The NVD reports contain several pieces of information, such as the affected software, the level of impact, the type of vulnerability, and the party responsible for disclosing the vulnerability
National Vulnerability Database

• To better assist the organizations and citizens of the Internet, the National Institute of Standards and Technology, with the support of the Department of Homeland Security, built the National Vulnerability Database (NVD)

• The NVD supplies all necessary information for security bulletins, including the affected software, level of impact, and the source

• Prior to the creation of the NVD there was no centralized location for storing all necessary information regarding vulnerabilities and exploits

• There were several databases each storing disparate information about the vulnerabilities and exploits, the NVD was built in an effort to create one centralized location for accessing all necessary information
Common Vulnerabilities and Exposures

• In 1999 the Common Vulnerabilities and Exposures (CVE) database was created in an effort to provide a central location for all vulnerability bulletins

• Prior to the creation of the CVE database each piece of software or website was responsible for sharing any vulnerabilities or exposures themselves,
  • Consumers were required to visit multiple software provider websites in order to ensure that they were not running software with known vulnerabilities

• The fields contained in each CVE entry are very minimal, only containing an ID, a description, the creation date, and a list of references
Common Platform Enumeration

• One of the aspects lacking from the CVE database is a well defined list of impacted software
• Other than parsing through the description or reading the reference links, there is no way to tell the exact application that the bulletin pertains to
• The CPE contains the identities of all software, operating systems, and even hardware that are used by a company or individual
• Provides a standardized method for organizing and identifying these elements so that they may more easily be recognized in a reliable way, especially programmatically, when referenced in various documents
Common Vulnerability Scoring System

• As a system administrator that is constantly bogged down with a variety of issues, both security related and not
• Can be very challenging to take the time to fully investigate every CVE bulletin posted to determine if it poses an immediate risk to your service
• The Common Vulnerability Scoring System (CVSS) was created in an effort to help vendors and domain experts to provide the context and impact level for each specific vulnerability described in the CVE database
Common Weakness Enumeration

• The Common Weakness Enumeration (CVE) was developed as a way to classify the specific weakness or type of attack that the CVE describes

• The CVE contains many hundreds of types of weaknesses, but the NVD only utilizes a handful for classifying the CVE bulletins that are contained in the NVD

• Some examples of CWE elements included in the NVD are: Authentication Issues, Buffer Errors, Buffer Errors, and SQL Injection
Trends in the NVD
Trends in the NVD
Trends in the NVD
Constructing the Heterogeneous Network

• The NVD data stream is essentially a merge of multiple different databases the obvious approach to creating a heterogeneous network is to simply separate out the databases into their disparate parts, creating the links between nodes as needed

• Five different data types were created and populated as the NVD data was read through

• The CVE identifier, the vulnerable software in the form of CPE identifiers, the base metric of the CVSS score, the weakness type as the CWE identifier, and the source identifier of the disclosure
Mining the Network

• The major outcome that the research conducted here wanted to accomplish was to find the CVE most similar to the Heartbleed CVE

• Goal was to attempt to eventually create a pre-warning signal that would find pieces of software that would likely have CVEs at the Heartbleed level so that research could theoretically be conducted specifically into those products in an effort to find the vulnerabilities before any nefarious individuals

• PathSim seemed like a good starting point for this path of vulnerability prediction
Mining the Network

• In 2011 Sun, et al, published a paper that pioneered the idea of utilizing the meta-path information from a heterogeneous network to find the top-k most similar objects to a query object

• Uses the example of mining paths in the DBLP database to find the most similar authors, and eventually builds to predicting which authors will become co-authors in the future

• Utilizes a symmetric meta-path, that is a path between two objects that are linked by common elements, to compute a similarity measure, which is then used to find the most similar items to the item in question
Future Work

• The meta-path is too vague to find anything interesting
• The other data types should be brought into the meta-path and examined
• Add in mining of the source fields in the NVD database
• The summary section of the NVD database elements could be mined to find valuable data on trends and predictions in vulnerabilities
• Compare the prediction abilities using the base CVSS score or using the other types of CVSS scores
• The CWE section can be explored in more detail
Conclusion

• The NVD provides valuable data on vulnerabilities and exposures
• This data can be used to construct a heterogeneous network
• The heterogeneous network can be mined with a version of PathSim to find interesting software predictions
• These predictions could help to keep the Internet safe
• Still a lot of future work to provide valuable predictions, but the foundation has been laid