Provsec: a Cybersecurity Incident Dataset for Hands-on Provenance Investigation Education and Research

Madhukar Shrestha, Yonghyun Kim, Jeehyun Oh

Faculty Advisors: Dr. Junghwan (John) Rhee, Dr. Fei Zuo, Dr. Grace Park

COMPUTER SCIENCE

UNIVERSITY OF Central Oklahoma

1 PROBLEM AND MOTIVATION

We are having a high number of cybersecurity incidents in most areas including industry, government, education, residential homes, and municipal agencies. Such attacks result in different malicious activities including data theft, DDoS, watering hole attacks, server hijacking, and many more.

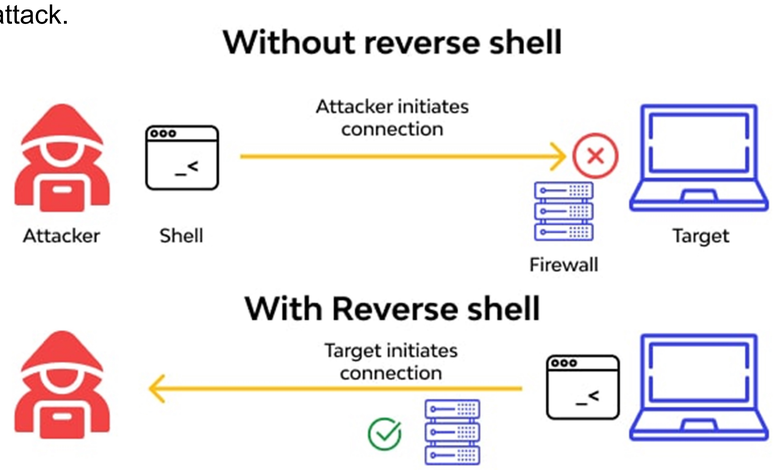
To respond to these cybercrimes, we need strong workforce called cybersecurity analysts who can investigate such attacks, recover infrastructures, and plan defense mechanisms for future attacks. However, a solid hands-on dataset for cybersecurity education and research is lacking.

2 BACKGROUND

Common Vulnerabilities and Exposures (CVE) are used to track and categorize vulnerabilities in consumer software. Such vulnerabilities are exploited by attackers to gain unauthorized access to the computer and/or network system.

Reverse shell can be achieved by exploiting the vulnerabilities whereby remote machine initiates the connection. This is an effective way to get a remote shell access over a NAT or firewall. Mostly incoming traffic are filtered by firewall but not for the outgoing traffic.

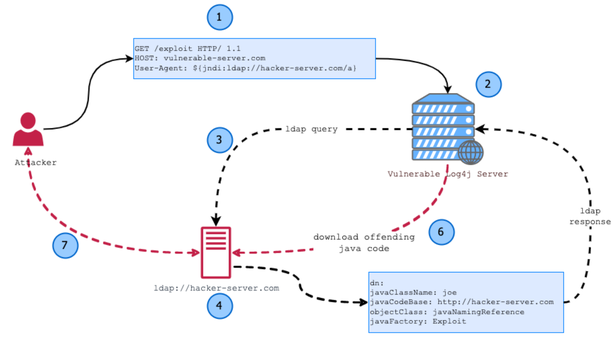
A backdoor attack occurs when threat actors create or use a backdoor to gain remote access to a system. Such attacks enable attackers to gain control of system resources, perform network reconnaissance and install different types of malware. In some cases, attackers design a worm or virus to take advantage of an existing backdoor created by the original developers or from an earlier attack.



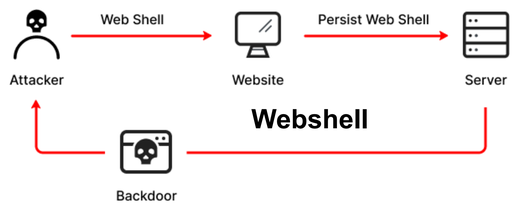
3 VULNERABILITIES AND EXPLOIT

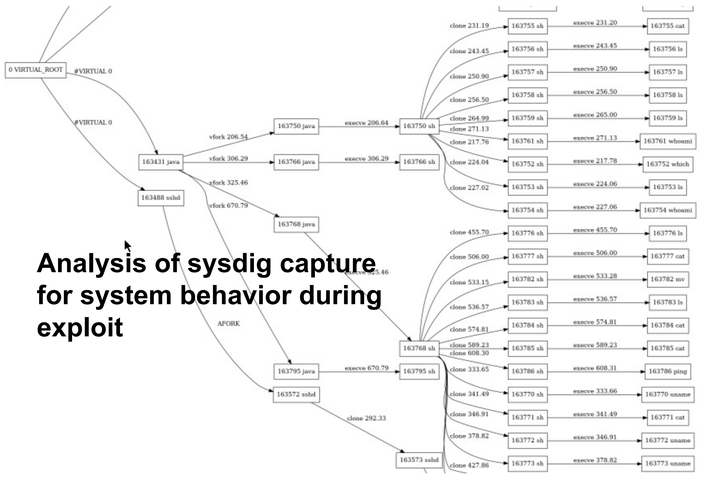
We explored different software vulnerabilities and performed the exploit for those CVEs. For each selected vulnerability, we prepared a safe sandbox environment to simulate the attack environment and captured the behavior, events and system calls during the exploit. Our dataset has the following characteristics:

- Container based application workload (Docker)
- Real vulnerability exploits for attacks
- Description on the ground truth of an attack for validation
- Two kinds of data (without and with an attack) useful for ML tasks
- Provides operating system events (e.g., system calls) to analyze the detailed system behavior.



Backdoor attack for CVE-2021-44228 5

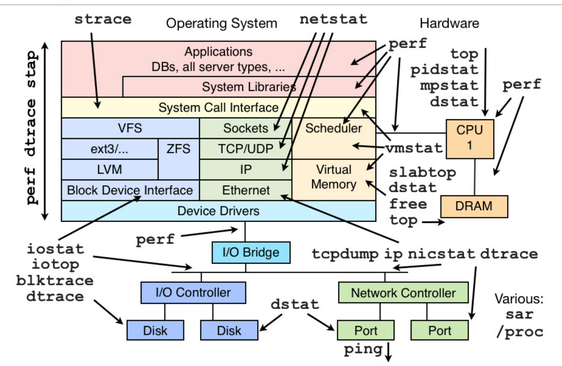




Analysis of sysdig capture for system behavior during exploit

4 ANALYSIS AND TOOLS

Sysdig is a Linux diagnostic tool used to capture system calls, libraries and network calls. It collects information on various resource usages of software:



Sysdig architecture

5 RESULT

We have built a dataset based on several well-known cases of cyber incidents and security vulnerabilities exploits. We analyzed the system behavior and libraries as well as network activities. This result helps to identify how an attacker can gain access to the system or execute arbitrary commands using different vulnerabilities. This dataset is currently contributing to research projects and coursework regarding cybersecurity incident analysis.

#	Software	Vulnerability	Target Process
1	nginx	CVE-2017-7529	nginx
2	apache	CVE-2021-41773	httpd
3	ghostscript	CVE-2018-16509	python
4	php	CVE-2018-19518	apache2
5	log4j	CVE-2021-44228	java
6	tomcat	CVE-2020-1938	java
7	redis	CVE-2022-0543	redis-server
8	consul	N/A	consul
9	apache	CVE-2021-42013	httpd
10	django	CVE-2021-35042	python
11	docker	CVE-2019-5736	docker