The Racing Game

A Study on Accelerating the Incident Response Cycle for Organizations
under attack in real time

Zezhou Wang, Zhengqing Ye, Liyun Li

Contents

1 Introduction

The current state of Internet Security remains a critical issue. A lot of publicly-exposed servers are still exploitable by vulnerabilities that were discovered many years ago. This research aims to correlate malicious network activities with potential attack vectors and affected parties in the United States. With the use of third-party data-gathering services predicting intrusion attempts against a certain organization in real-time becomes feasible. Throughout the document, core concepts and technical details and methodologies is discussed thoroughly.

2 Acronyms

This section provides a list of acronyms used throughout the documentation.

- AWS: Amazon Web Service
- CVE: Common Vulnerabilities and Exposures
- IoT: Internet of Things
- IP: Internet Protocol
- ISP: Internet Service Provider
- JSON: JavaScript Object Notation
- KNN: K-Nearest Neighbors
- NLP: Natrual Language Processing
- OS: Operating System
- SCADA: Supervisory Control and Data Acquisition
- SQL: Structured Query Language
- SVM: Support Vector Machine
- US: United States

3 Background

This section describes the motivation and decisions throughout the research. Initially, the research was conducted to discover the cause of unauthorized network intrusion attempts in the world to understand attackers better. However, due to time constraints and a lack of resources, the priority is to analyze traffic within the United States, by performing infrastructure vulnerability research on servers predicted to be under attack.

3.1 Motivation

The intent is to accelerate the reporting of incident response to an organization if our automation detects a known vulnerability in their infrastructure so that the responsible team is able to realize and mitigate the threat more quickly.

In case the explanation does not clarify the motivation, an analogy is used among current and prospective researchers for understanding the motivation: The project is like a campaign to solve world hunger. It is very difficult to launch a campaign that can raise sufficient funding to feed every hungry person, but with time, more and more people will receive aid if they are within our reach.

3.2 Limitation

Since the research is conducted based on calculating the probability that the approximate coordinate of a victim is accurate by the gazetteer data set, the gazetteer must contain sufficient amount of data. If a location is under attack but its geographical location is not present in the database, the victim cannot be identified. Hence the scavenging of gazetteer must continue to a point where the automation is confident about accurately locating the victim within the US soil.

3.3 Tooling

The project uses Shodan for fingerprinting infrastructures, gazetteer lists for pinpointing victim location and various threat maps as a light reference for monitoring real-time intrusions.

3.3.1 Shodan

The search engine for the IoTs can help provide a novel data source for information gathering. Shodan is used for the following purposes:

- Crawling and indexing billions of devices in the US.
- Using various search filters to find indexed devices including SCADA devices.
- Retrieving network (port, IP), location (latitude, longitude), and fingerprints (banner, protocol, host names, etc.) pertaining to the devices.

Shodan's stream API provides a great understanding on device exposure. It should be noted that previous work on the Shodan database has shown that devices are generally indexed within 3 weeks of coming online (Ercolani, 2016). Although additional work done on evaluating Shodan as a scanning tool does show inconsistency in results from scans, analysis of the data from the scans has not been delved into for deeper analysis.

3.3.2 Gazetteers

At the moment, the source of the gazetteers are from the United States Census Bureau's collection in 2018. Files from the past is not included since the trainer must receive the most up-to-date data.

3.3.3 Threat Maps

Although companies claim that their threat maps are real-time, it might not be the case for threat detection. Additionally, due to the anonymous nature of the Internet, locations provided by threat maps cannot be trusted. For example, a lot of devices in rural areas are mapped to a single ISP node with the same latitude and longitude even though it can cover over 400 square miles. Therefore threat maps in general can only be used for light reference to selectively assess a target.

4 Pipeline Workflow

Below is an visual illustration of the workflow.

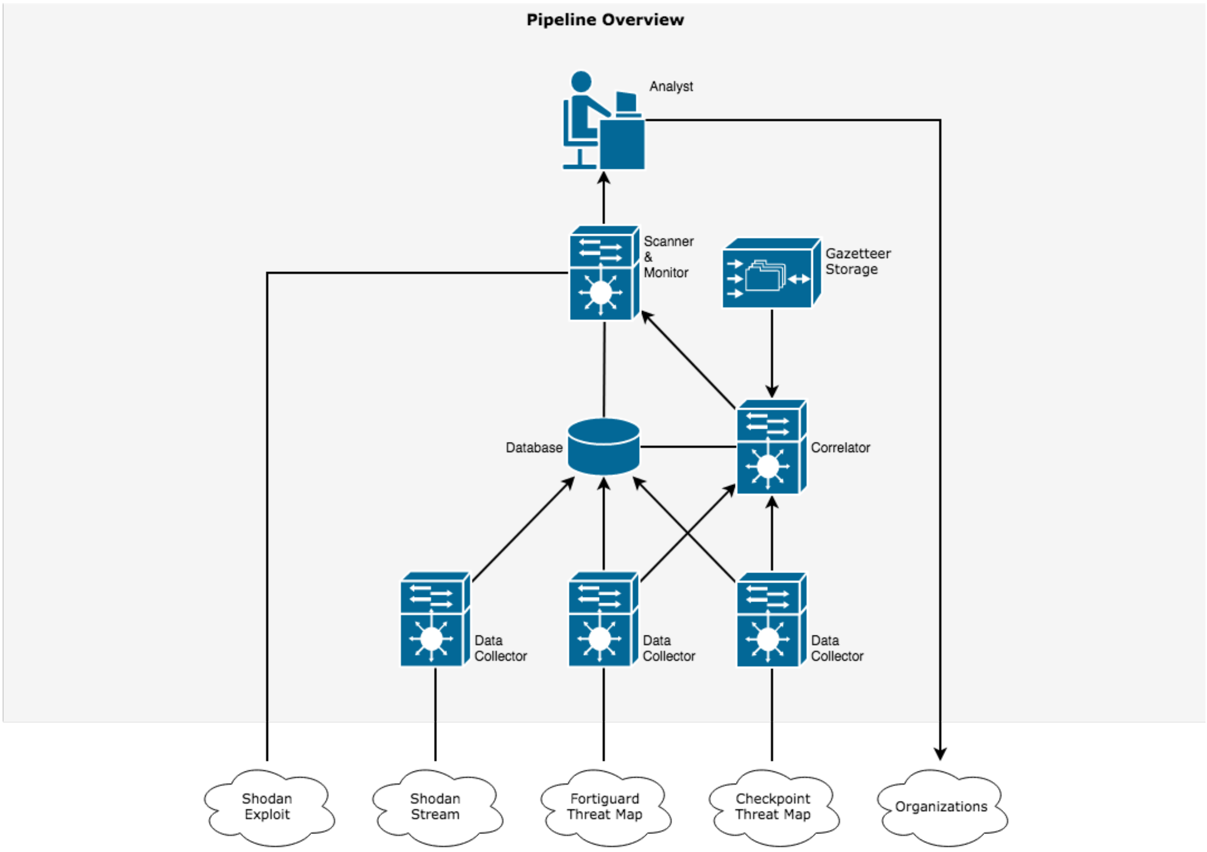


Figure 1: Pipeline Workflow

4.1 External Entities

The external entities are the cloud-shaped entities outside of the control of the pipeline.

- The Shodan exploit entity is an API that queries for know CVEs and exploits given a search string. It uses the https://exploit-db.com as its source for exploits.
- The Shodan stream is an API endpoint that feeds in server banners
- The threat map entities are websockets that yield real-time threat traffic

- The Organizations are organizations that is under attack in real time.

4.2 Internal Entities

The internal entities are objects within the grey area, an indication of controllable variables.

- Data collectors are essentially automated Python programs that continuously collect and parse data streams.
- The database holds all the parsed stream data. In our implementation it is a PostgreSQL instance.
- The correlator trains the gazetteers and the threat maps, and then correlate them with a Shodan entity in the database. If it is likely that an attach is directed towards a server in the database, it reports the traffic to Scanner.
- The Scanner & Monitor monitors for real-time attack, retrieves the fingerprints of the victim in question, queries the Shodan Exploit database.
- Once an attack is identified, the analyst verifies the attack and reports it to the organization.

5 Data Set Discussion

In order to retrieve real-time data, we use the Shodan stream API. Two queries are used for the project, the banner query and the exploitation query. The banner query returns information about servers in JSON format. It contains the following useful properties for the purpose of research:

5.1 Main Variables

This is the list of properties required by the research.

- IP addresses
- Port number
- Time stamp
- Host name
- Latitude and longitude
- The organization with the assigned IP space
- ISP, OS and transport method

5.2 Additional Variables

The following items serve as additional entries to aid the research:

- Title of website within the HTML source
- Name and version of the product
- Device type
- Relevant Common Platform Enumeration of product

For the exploit query, it can information about the source of the vulnerability and platform and port that is associated with this vulnerability.

6 Analysis Methodology

The items listed below is strictly followed to actualize the flow. The Shodan API flow is not included because it should be a continuous process.

1. Intercept threats from various threat maps and approximate the longitude and latitude and only locate US destinations.
2. Direct the approximated coordinates to the analyzer which attempts to map them to the gazetteer for a collection of geolocations.
3. If there is a hit, query the Shodan stream database for the coordinates and extract host information.
4. Perform passive fingerprinting (because active fingerprinting requires permissions from the organization in question) to predict the attack surface and validate the vulnerabilities.
5. If the mapping attempts in any of the steps above fails, queue the query to run again in the future.

7 Current Analysis and Results

This section documents the completion status and current progress.

7.1 Completed

The Shodan The data gathering of gazetteers should be sufficient to successfully locate at least one intrusion attempt within a day. We are able to selectively perform passive scanning for an organization suspected to be under siege, which essentially contributes to fulfilling the research purpose by informing the organization about potential breach.

Statistics are generated to further increase the probability that the victim is in fact the victim. For example, the graph below shows the most used attacks of all time.

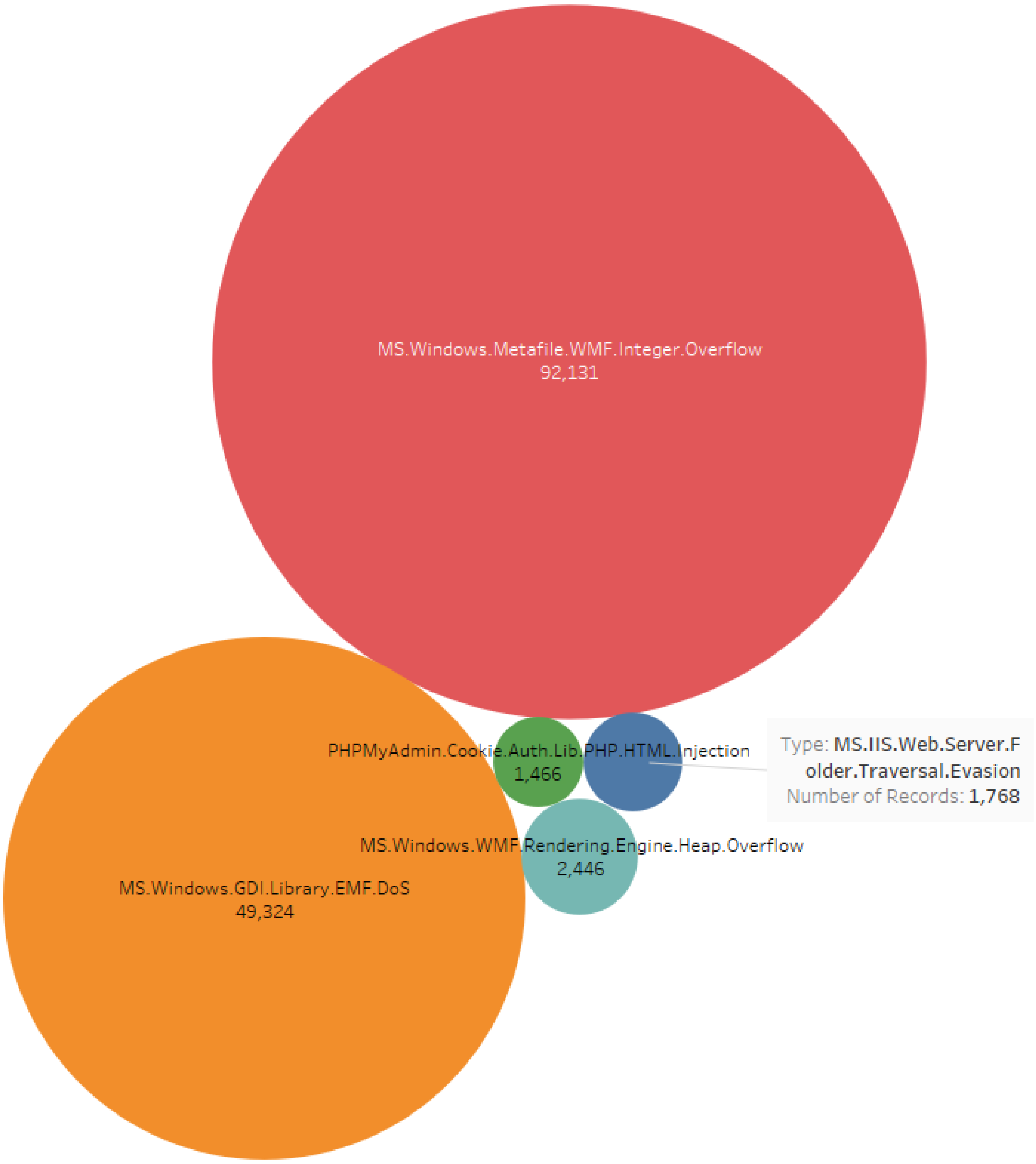
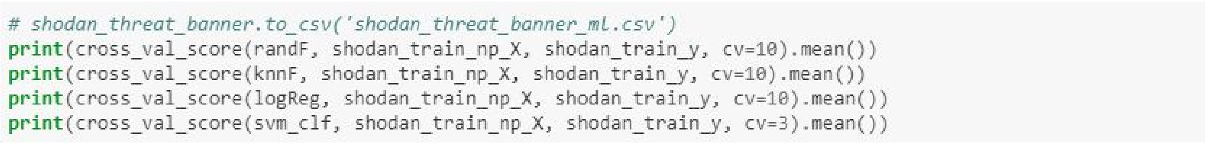


Figure 2: Top attacks all-time

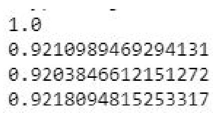
Unfortunately, the gazetteer data set are not large enough for detecting most of the attacks as the banners from Shodan are usually not in the keyword identifiers, making it extremely difficult to get a good and reliable F1 score. A workaround is temporarily in place which is pinpoint the targets.

We have compared several machine learning algorithms including SVM, Random Forest and Linear Regression. The result indicated that the Random Forest model is best suited for our purposes. Note that there may be discrepency here as the data set might not be large enough. It could also be because the data set is overfitting the model. If that is the case, the SVM is the winner here.



```
# shodan_threat_banner.to_csv('shodan_threat_banner_ml.csv')
print(cross_val_score(randF, shodan_train_np_X, shodan_train_y, cv=10).mean())
print(cross_val_score(knnF, shodan_train_np_X, shodan_train_y, cv=10).mean())
print(cross_val_score(logReg, shodan_train_np_X, shodan_train_y, cv=10).mean())
print(cross_val_score(svm_clf, shodan_train_np_X, shodan_train_y, cv=3).mean())
```

Figure 3: Snippet of what is being printed



```
1.0
0.9210989469294131
0.9203846612151272
0.9218094815253317
```

Figure 4: Output

7.2 In Progress

Hence a huge amount of effort is put in the scavenge hunt for gazetteer. The historical data from the Census Bureau must be further analyzed before it is passed to the trainer.

In addition, the machine learning model can only train a set of data. In order to analyze an incoming record, an NLP model is required so more efforts must be put in that.

8 References

- https://github.com/liyun-li/security-analytics-project