Vulnerability Analysis of Internet Devices from Indonesia Based on Exposure Data in Shodan

B Novianto*, Y Suryanto and K Ramli
Department of Electrical Engineering, Faculty of Engineering, University of Indonesia, Depok, Indonesia
*Corresponding author: bambang.novianto81@ui.ac.id

Abstract. The growth of internet-enabled devices has increased interest in cybersecurity. In 2014, Project SHINE (SHodan INtelligence Extraction) published a report of large-scale security assessments for devices connected to the Internet. However, the number of IP addresses harvested from Indonesia in 2014 is very small. There were 7,182 IP address from Indonesia. It was about 0.0032% from the total 2,186,971 IP addresses. In this paper, we propose an initiative to gather all information for all Autonomous System Number (AS Number) from Indonesia in Shodan. We have gathered a dataset about all information of AS Numbers in Indonesia such as 12,787 unique ports, 79 unique operating systems, 409 unique products, 3,634 unique domains, 145,543 unique IP addresses, and 790 unique organizations. We use the K-Means algorithm to cluster all AS Numbers into several classes according to the exposure level in shodan. Based on the result, we have 4 classes of AS Numbers. There are 1,075 AS Numbers in class:0 (no information in Shodan yet), 614 AS Numbers in class:1 (exposure level = low), 9 AS Numbers in class:2 (exposure level = medium), and 1 AS Number in class:3 (exposure level = high). This information can be used to warn the organizations that manage AS Numbers in Indonesia to be aware of the security and the threats to their systems.

1. Introduction
Nowadays, the Internet has been one of the most transformative and fast-growing technologies. The growth of internet-enabled devices has increased interest in the security of data. To collect data from internet-enabled devices, Project SHINE was created. Project SHINE collects data by using the Shodan search engine. Shodan is a search engine that scans the internet to find open ports. Shodan is a reference to “Sentient Hyper-Optimized Data Access Network” [1]. Shodan’s information is considered valuable when attempting to discover critical systems that are directly connected to the Internet. There will be a positive or negative aspect of Shodan. It depends on its user’s intent. Shodan can be used as either an assessment portal or as a weapon for hackers to scan targets passively.

Based on Project SHINE reports in 2014, there were 211 countries identified based on the IP addresses harvested. The highest countries are the United States with 616,994 IP addresses (33.75%), Germany with 280,248 IP addresses (15.33%), China with 112,114 IP addresses (6.13%), Korea with 99,856 IP addresses (5.46%) and United Kingdom with 66,234 IP addresses (3.62%). Unfortunately, the number of IP addresses harvested from Indonesia in 2014 is very small. There are 7,182 IP address (0.0032% from the total 2,186,971 IP addresses).

In this paper, we propose an initiative to gather all information for all Autonomous System Numbers (AS Numbers) in Indonesia from Shodan. An autonomous system is a group of connected Internet Protocol under the control from a single administrative entity. We use a clustering algorithm to
categorize all AS Numbers in Indonesia. Clustering is one of the most widely studied problems in data mining [2]. Clustering is the process of grouping a set of unlabeled objects into classes whose members are similar in some way [3]. There are many methods for clustering. The most common method is partitioning-based clustering. This method divides the large dataset into a number K of groups by using a distance as a measure of similarity, where each group represents a cluster [4]. K-Means [5], K-Medoids [6], K-Modes [7], PAM [8], CLARA [9], CLARANS [10] are the most algorithms that use this method. We use K-Means clustering because it is the most popular distance-based partitional clustering algorithm that is simple yet efficient [11]. The result of this study can be used to warn all organizations in Indonesia that manage AS Number to be aware of potential vulnerabilities that can be exposed from Shodan.

The structure of the rest of this paper is as follows. We provide related works in Section II. In Section III, we show the methodology that we used. Section IV, we present the result and discussion of this research. In Section V, we present the conclusion.

2. Related Works

Our first reference is the SHINE project [1]. SHINE project conducted a large-scale security assessment for devices connected to the Internet. There were 211 countries identified based on the IP addresses discovered. The highest country was the United States with 616,994 IP address (33.75%). The number of IP addresses harvested from Indonesia in 2014 is very small. There were 7182 IP address only. It is about 0.0032% from the total 2,186,971 IP address discovered.

Ercolani et al. [12] visualized IP addresses and open ports from Shodan. The dataset was derived from the set of all scans from the 1st through the 14th of September 2015. They extracted all IP addresses that were identified from Iran. Ali et al. [13] used Shodan to present vulnerability scanning for IoT devices in Jordan. They wanted to warn the community about IoT security issues and could be exposed. They pointed-out the critical cyber-attack against IoT and presented statistics about vulnerable devices in Jordan. Rae et al. [14] surveyed methods of assessment using Shodan. They presented the impact of Shodan to identify Internet-connected devices for both good and bad purposes. Williams et al. [15] assessed IoT devices exposed on the Internet with Shodan. However, their focus were on the vulnerabilities of webcams, smart TV, and printers. They collected 8,127 data of Webcams, 1,437 data of Smart TV, and 10,675 data of Printers.

There are many different kinds of research to assess internet-enabled devices with Shodan. However, none that focus to analyze and gather all information for all AS Numbers from Indonesia in Shodan. Ercolani et al. [12] and Ali et al. [13] did their research in their own countries such as Iran and Jordan. SHINE project [1] did a global-scale assessment, but they were not focused on Indonesia. The number of IP addresses from Indonesia harvested by the SHINE project is very small. Williams et al. [15] just focused their assessment on webcams, smart TV, and printers.

3. Methodology

The main purpose to study the exposure of all internet-enabled devices in Shodan is to be able to profile vulnerabilities in Indonesia. Because there are a huge number of IP address from Indonesia, we cluster all IP address based on the AS Numbers. These are our methodology for this study:

3.1. Data Collection

In this step, we will create a dataset by collecting all information in Shodan for all AS Numbers from Indonesia. First of all, we collect all AS Numbers in Indonesia. We can get information about the list of AS Numbers in Indonesia from ipinfo.io [16]. Then, we gather all subnets for each AS Numbers. We can look up all subnets for each AS Number by using API from hackertarget.com. We create bash scripting to collect all subnet for each AS Number into CSV files. The fields of the data are as follow.

| Fields       | Information                  |
|--------------|------------------------------|
| AS Number    | Autonomous System Number     |
| Subnet       | List of subnets for each AS Number |
We scan all subnets for each AS Number in Shodan. We register a business account to get the API Key of Shodan. We create bash scripting to automatically scan all subnets for each AS Number from Indonesia in Shodan. Daily downloads of scans have been parsed and inserted into CSV files. The dataset was derived from December 2019 until January 2020. We observed 1,699 AS Numbers in Indonesia. These are the fields that we collect from Shodan.

| Fields         | Information                                           |
|----------------|-------------------------------------------------------|
| data           | Information with raw data                             |
| domains        | Information about Domain Name System                  |
| hash           | Hash value of information                             |
| host           | Information about host or user                        |
| hostnames      | Information about hostnames                           |
| info           | Additional information                                |
| ip_str         | IP address of the target                              |
| isp            | Information about Internet Service Provider           |
| org            | Information about the organization                    |
| os             | Information about operating system                    |
| port           | Information about port                                |
| product        | Information about product of services                 |
| tags           | Tags of information                                   |
| timestamp      | Timestamp of information                              |
| transport      | Information about protocol transport                  |
| version        | Information about the version of services             |

### Table 2. The fields of data from shodan

3.2. Data Cleaning and Integration

We use Splunk Data Analytics for indexing data in a searchable repository that can generate reports, dashboards, and visualizations. Before we input our dataset into Splunk, we need to clean our data from blank spaces. We integrate all CSV files for each AS Number in Shodan into a single file CSV. Then, we convert semicolon CSV into a comma-delimited CSV for input to Splunk.

3.3. Data Featuring

For the efficiency of processing, we need to keep the number of feature space as low as possible but should present host behavior carrying rich enough information. We found the following features are the most important to represent the exposure quantity in shodan:

- **Number of queries**
  
  This feature represents the number of queries for each AS Number in Shodan. If there are many queries that we collect from Shodan, it indicates that the exposure level for that AS Number is high. If there are few queries that we collect from Shodan, it indicates that the exposure level for that AS Number is low.

- **Size (in bytes)**
  
  This feature represents the numbers of information in bytes. If the size of the data that we collect from Shodan is big, it indicates that the exposure level for that AS Number is high. If the size of the data that we collect from Shodan is small, it indicates that the exposure level for that AS Number is low.
3.4. Data Mining

We use K-Means Algorithm to cluster all AS Numbers based on the level of exposure fin Shodan. K-Means Algorithm is the most commonly used clustering algorithm. The distance measure it generally uses is Euclidean distance, given by:

\[ \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \]  \hspace{1cm} (1)

where \( x_i \) and \( y_i \) are two points in a d-dimensional Euclidean space. The objective function to be minimized is called Sum of Squared Error (SSE), given by the following equation:

\[ SSE = \sum_{i=1}^{d} \sum_{x \in c_k} (x_i - c_k)^2 \]  \hspace{1cm} (2)

The cluster centroid \( c_k \) can be updated as:

\[ c_k = \frac{\sum_{x \in c_k} x_i}{c_k} \]  \hspace{1cm} (3)

Here is the algorithm of K-Means clustering [17].

---

### K-Means Clustering Algorithm

**Input:** \( k \) (number of clusters), \( D \) (dataset)

**Output:** a set of \( k \) clusters

**Method:**
- Arbitrarily choose \( k \) objects from dataset \( D \) as an initial cluster;

**Repeat:**
- 1: reassign each object to the cluster that most similar based on the mean value in the cluster
- 2: update the cluster means by calculating the means value of the objects in the cluster

**Until** no update

---

3.5. Evaluation

We evaluate vulnerabilities of AS Numbers from Indonesia based on the exposure level in Shodan. In this paper, we provide several visualization and statistics about ports, products/services, operating systems, and domains. Information detail about the name of organizations and all IP address that exposed in Shodan is not presented in this paper to protect the security of the organizations and IP address.

4. Numeric Examples

We observe 1.699 AS Numbers in Indonesia. There are 624 AS Numbers information that we collect from Shodan such as IP address, ports, product services, domains, operating systems, and organizations. We gather 289,715 queries of Shodan into a single file. The size of the file is 237 MB. These are all the information that we collect from Shodan.

4.1. Ports

There are 272,457 information about ports. From 272,457 information, there are 12,787 unique ports. Port 80 is the highest port from Indonesia in our dataset. There are 380,008 information about port 80 (13.95%). Figure 1 provides us information about the statistics about ports from Indonesia in Shodan.
There is 2.141 information about operating systems. From 2.141 information, there are 79 unique operating systems. Linux 3.0 is the highest operating system in this dataset. There are 768 information about Linux 3.x (35.871%). Figure 2 provides us information about the statistics of operating systems from Indonesia in Shodan.

4.3 Product/Services

There are 98.152 information about product services. From 98.152 information, there are 409 unique products. “MikroTik bandwidth-test server” is the highest product in this dataset. The information about product/services are MikroTik bandwidth-test server (32.759 information), apache httpd (16.236 information), Open SSH (7.638 information), nginx (5.183 information), ntpd (3.343 information), Mikrotik router ftpd, (3.118 information), Dropbear sshd (3.088 information), MySQL (2.456 information), Postfix smtpd (2.391 information), etc. Figure 3 provides us information about the statistics of products/services from Indonesia in Shodan.
Figure 3. Statistics of products/services from Indonesia in Shodan

4.4 Domains

There are 111,732 information about domains. From 111,732 information, there are 3,350 unique domains. The highest top-level domain is *.id. Table 3 gives us information about the Top-Level Domains from Indonesia in Shodan.

Table 3. Top-Level Domains from Indonesia in Shodan

| Top-Level Domains | Number of Unique Domains | Total Information |
|-------------------|--------------------------|-------------------|
| *.id              | 1,881                    | 85,238            |
| *.com             | 1,242                    | 17,498            |
| *.net             | 190                      | 8,190             |
| *.arpa            | 15                       | 666               |
| *.tv              | 9                        | 91                |
| *.biz             | 13                       | 49                |

4.5 IP addresses

There are 272,461 information about IP addresses. From 272,461 information, there are 145,543 unique IP addresses. Unfortunately, for security and secrecy reasons, we cannot publish all specific IP addresses in this paper because the information will refer to specific organizations in Indonesia.

4.6 Organizations

There are 271,731 information that includes the name of the organizations such as Internet Service Providers, agencies, and universities. From 271,731 information, there are 790 unique organizations. Unfortunately, we can not publish the name of organizations in this paper for security and secrecy reasons. The highest organizations in Shodan has 64,505 information (23,738%).
4.7 Clustering for AS Number in Indonesia

There are 1.699 AS Numbers in Indonesia. We create a dataset about all information of 624 AS Numbers that exposed in Shodan. There are 1.075 AS Numbers that are not exposed to Shodan. 1.075 AS Numbers will be categorized as class:0 that indicates that there is no information in Shodan yet. According to the number of information in Shodan, we will cluster 624 AS Numbers to be several classes with the K-Means algorithm. We use two features for the K-Means algorithm. The features are the “number of the query” and “size (in bytes)”. These features represent the quantity of information that we collect from Shodan including ports, IP addresses, products, operating systems, domains, and organizations. We want to cluster the exposure level to be 3 classes such as: low, medium and high. These are the result of the K-Means algorithm with value of centroid K=3.

Figure 4. The result of K-Means clustering algorithm

Figure 4 shows us that there are many blue nodes (614 nodes) as class:1, several green nodes (9 nodes) as class:2 and a single orange node as class:3. We have tried for centroid K>3. The result is not better than centroid K=3 because there is no clear boundary for the blue nodes to be separated into more classes.

5. Conclusion

We successfully create a dataset about all information of AS Numbers from Indonesia in Shodan. From our dataset, there are 272,457 information about ports (with 12,787 unique ports), 2,141 information about operating systems (with 79 unique operating system), 98,152 information about product services (with 409 unique products), 113,358 information about domains (with 3,634 unique domains), 272,461 information about IP addresses (with 145,543 unique IP addresses), and 271,731 information about the name of organization such as Internet Service Providers, agencies, and universities (790 unique organizations). Based on the result of the clustering algorithm, we get four classes of AS Number related to the level of exposure in Shodan such as:
• Class: 0.
  This class indicates that there is no information about the AS Number in Shodan yet. There are 1,075 AS Numbers in this class.
• Class: 1.
  This class indicates that there is low information about AS Number in Shodan. There are 614 AS Numbers in this class.
• Class: 2.
  This class indicates that there is medium information about AS Number in Shodan. There are 9 AS Numbers in this class.
• Class: 3:
  This class indicates that there is high information about AS Number in Shodan. There is 1 AS Number in this class.

This information is important to warn organizations that manage the AS Number to be aware of the security of their systems.

Acknowledgments

This paper’s publication is supported by Hibah PITTA UI 2020 funded by the University of Indonesia.

References

[1] Koblenz G. D 2020 Emerging technologies and the future of CBRN terrorism. The Washington Quarterly 43 (2) 177-196
[2] Chehouri A, Younes R, Khoder J, Perron J and Ilinca A 2017 A selection process for genetic algorithm using clustering analysis. Algorithms 10 (4) 123
[3] Jain A K, Murty M N and Flynn P J 1999 Data clustering: a review ACM computing surveys (CSUR) 31 (3) 264-323
[4] Kumar K M and Reddy A R M 2016 A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method Pattern Recognition 58 39-48
[5] Liu X, Zhu X, Li M, Wang L, Zhu E, Liu T and Gao W 2019 Multiple kernel $ k $-means with incomplete kernels IEEE transactions on pattern analysis and machine intelligence 42 (5) 1191-1204
[6] Sammut C and Webb G I 2017 Encyclopedia of machine learning and data mining Springer
[7] Jiang F, Liu G, Du J and Sui Y 2016 Initialization of K-modes clustering using outlier detection techniques Information Sciences 332 167-183
[8] Ng R T and Han J 2002 CLARANS: A method for clustering objects for spatial data mining IEEE transactions on knowledge and data engineering 14 (5) 1003-1016
[9] Gupta T and Panda S P 2018 A comparison of k-means clustering algorithm and clara clustering algorithm on iris dataset International Journal of Engineering & Technology 7 (4) 4766-4768
[10] Matioli L C, Santos S R, Kleina M and Leite E A 2018 A new algorithm for clustering based on kernel density estimation Journal of Applied Statistics 45 (2) 347-366
[11] Jain A K 2010 Data clustering: 50 years beyond K-means Pattern recognition letters 31 (8) 651-666
[12] Fernández-Caramés T M and Fraga-Lamas P 2020 Teaching and learning iot cybersecurity and vulnerability assessment with shodan through practical use cases Sensors 20 (11) 3048
[13] Yu M, Zhuge J, Cao M, Shi Z and Jiang L 2020 A survey of security vulnerability analysis discovery detection and mitigation on IoT devices Future Internet 12 (2) 27
[14] Darlaston J and Wintle J 2007 Safety factors in the design and use of pressure equipment Engineering failure analysis 14 (3) 471-480
[15] Alshehri M D, Hussain F K and Hussain O K 2018 Clustering-driven intelligent trust management methodology for the internet of things (CITM-IoT) Mobile networks and applications 23 (3) 419-431
[16] Rubner Y, Puzicha J, Tomasi C and Buhmann J M 2001 Empirical evaluation of dissimilarity measures for color and texture Computer vision and image understanding 84 (1) 25-43