Visualization of Encrypted Data Packages using Birch Algorithm

Herri Setiawan, Ahmad Sanmarino, Dwi Asa Verano

Abstract: Encrypted data packages are already widely used in public services; such as Secure Sockets Layer (SSL/Transport Layer Security (TLS) and tunneling. It is used for its advantage of providing secure data exchange. However, there are difficulties to recognize encrypted packet patterns and distinguish between safe and malicious data. This paper proposes a method to visualize encrypted data packages in graphical form using Birch algorithm where itrecognizes better.There are four stages carried out in the study, namely System Design, Feature Extraction, Cluster Process, and Visualization. The results of clustering using the birch algorithm of 0.98 or have a percentage level of accuracy of 98 percent.

Keywords: Encrypted,Cluster, Data Packages, Birch, Visualize

I. INTRODUCTION

The data transmission process requires security and trust to ensure data is safe from the interference of any forms. An increasing number of services on the internetsuch as posting and search have used applications based on SSL/TLS protocol to replace hypertext based services[1]. This is reinforced by research conducted by [2]which states that 95.4% of public services have already use the SSL/TLS protocol.

One method of hierarchical clustering is Birch[3]. There are two basic concepts in this method, namely clustering feature and clustering feature tree (CF tree), this method is able to perform a fast and good clustering process in large database clustering processes, this is because the structures of this method summarize cluster process.In addition, this method can summarize existing subcluster to reduce the scale of clustering. Whereas according to [4] the superiority of this method is that it can overcome difficulties in terms of scalability and can overcome the other algorithm's inability to repeat and cancel the previous execution.

Visualization will be better if combined with data mining and machine learning techniques in visualizing a data package. Visualization is a process that relates continuously between visualization and the knowledge discovery that functions in terms of data collection. Research by [5] visualizes attacks by displaying two-dimensional graphs with several stages, namely preparation, clustering, and visualization.By doing clustering, normal and abnormal data packages can be distinguished[6]

The problem in this research is how to do the process of grouping data packages with cluster methods and visualizing them in the form of graphs of data packet forms. The stages in this study are Capture Packet, Feature Extraction, and Features Selection. While the purpose of this study is to be able to recognize the patterns of encrypted packages and also be able to display the results in graphical form.

In this study will be discussed about the clustering process and the visualization of encrypted data packages, the theoretical basis will be discussed in part II, the methodological stages used in solving the problems are explained in section III and part IV which show the results. The conclusion will be shown in Section V

II. LITERATURE REVIEW

In a study conducted by [7] mentioning that the implementation of SSL was first developed by Netscape Communications Corporation in the early 1990s, it was intended to secure HTTP services. Then the Internet Engineering Task Force and Netscape started to build SSL 3.0 and in 2000 started standardization for SSL known as TLS. Fig. 1 describes the SSL Handshake protocol.

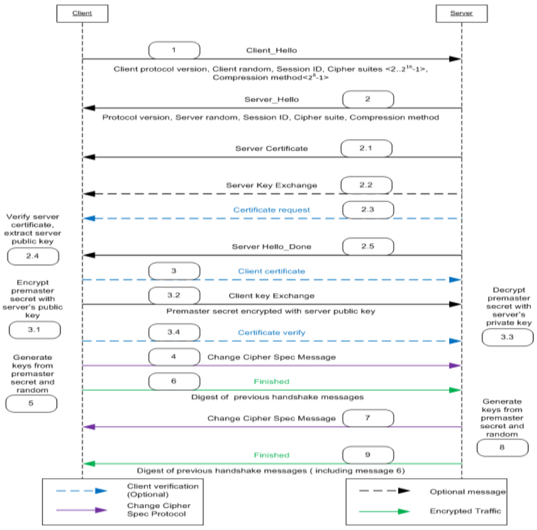
Client
Server
1
Client_Hello
Client protocol version, Client random, Session ID, Cipher suites <2..2^16-1>, Compression method<2^8-1>
Server_Hello
2
Protocol version, Server random, Session ID, Cipher suite, Compression method
Server Certificate
2.1
Server Key Exchange
2.2
Certificate request
2.3
Verify server certificate, extract server public key
2.4
Server Hello_Done
2.5
3
Client certificate
Encrypt premaster secret with server's public key
3.1
3.2
Client key Exchange
Premaster secret encrypted with server public key
Decrypt premaster secret with server's private key
3.3
3.4
Certificate verify
4
Change Cipher Spec Message
Generate keys from premaster secret and random
5
6
Finished
Digest of previous handshake messages
Change Cipher Spec Message
7
Generate keys from premaster secret and random
8
Finished
9
Digest of previous handshake messages (including message 6)
Client verification (Optional)
Change Cipher Spec Protocol
Optional message
Encrypted Traffic

Fig. 1SSL Handshake Protocol[8]

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) has adjacent data properties to be subcluster, meaning data that has a decent correlation will be grouping into a small cluster shape. This method uses the Cluster Feature (CF) function that concludes subcluster-subcluster into height balance tree.

Revised Manuscript Received on September 22, 2019.

Herri Setiawan, Department of Informatics, Faculty of Computer Science, Universitas Indo Global Mandiri

Ahmad Sanmarino, Department of InformationSystem, Faculty of Computer Science, Universitas Indo Global Mandiri

DwiAsaVerano, Master of EngineeringInformatics t of Computer System, Department, Universitas Sriwijaya

Retrieval Number A2693109119/2019©BEIESP
DOI: 10.35940/ijeat.A2693.109119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

N is the dimension of the data point in a cluster: {i}, where i = 1, 2, ... N, then the centroid value can be determined by the equation below

$$\overrightarrow{X0} = \sum_{i=1}^{N} \overrightarrow{Xi} \qquad (1)$$

The equation for calculating the average distance of members to points is

$$R = \left(\frac{\sum_{i=1}^{N} (\overrightarrow{X0} - \overrightarrow{X1})^2}{N}\right)^{1/2} \qquad (2)$$

And the equation for the average pairing distance in a cluster is:

$$R = \left(\frac{\sum_{i=1}^{N} \quad \sum_{j=1}^{N} (\overrightarrow{X1} - \overrightarrow{Xj})^2}{N(N-1)}\right)^{1/2} \qquad (3)$$

One way to get repetitive data between visualization and knowledge discovery is by visualization methods. research made by[9], [10]displays raw data with several parameters in two-dimensional graphs using parallel coordinates

III. RESEARCH METHOD

Fig. 2 is the roadmap of this study. There are four stages to be carried out, where each stage will be completed during the research process.

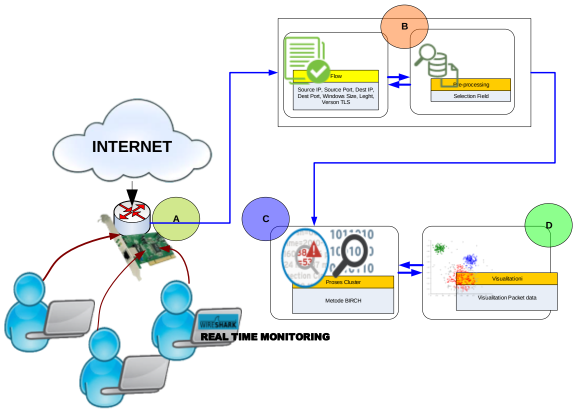

Fig.2 Research Roadmap

A. System Design

In this study, traffic monitoring will be centered on the Indo Global Mandiri Palembang Server (VPS) campus. This study uses several tools such as routers that function as data packet forwarders. Switches are used to share access to users and 1 PC that are used as monitoring and capture. real-time

B. FeatureExtraction

In this study, the features extraction process was built with an algorithm to extract raw data, aimed at obtaining values from attributes. The results of the features extraction process from raw data will produce a .csv file, where this file type is easy to process and generally accepted.

Strong attributes for the resulting are used as patterns in recognizing a package. In this study, the selection feature is used to find influential features and override features that have no effect

C. Cluster Process

The purpose of this process is to get the results of grouping encrypted data packages into clusters, in this experiment will produce two TLS v1.0 clusters and TLS V1.2. In the clustering experiment, it was determined that the initial centroid was a pattern of the encrypted data package.

D. Visualization

The final stage in the research is to visualize the forms of groups of data packets produced in the graphic form.

The dataset taken from the results of extracting data on the network will be done by the clustering process using the BIRCH algorithm. The processed data will be divided into three clusters. Each cluster will have a certain data pattern.

Procedure for clustering Data

```
Input : dataset (csv file)
Output : cluster BIRCH Result.
Def processing-data (filename) do
For 1 in range (Len input) do
Data
convert□ i[ttl],i[window],i[p_dest],i[flags],i[total-
length],i[protocol]
Write (data convert file-name) Return file-name
end
def BIRCH (data file result cluster ) do import
clustering Birch
read □(data file) Birch calculate □ birch (read,3)
Birch-calculate.process() C□ Birchcalculate.get
cluster() Write (c, result cluster)
Visualizer □ cluster_visualizer (c)
End
```

IV. RESULTS

The results of reading raw data strings combined with reading TCP Headers and UDP header is obtained by the number of packets of five tests as presented in table 1, and visualization of the encrypted packet data shown in Fig. 3.The test was conducted five times, with time functioning as an interrupt.

Table 1 is the result of the feature extraction process from the raw data obtained, it can be seen that the number of results from the feature extraction process is different from the number of raw data packets, this is because the applications built in this process only use TCP and UDP communication protocols that are processed (filter).

Table.1 Results of processing encrypted data packages

Testing	TLS V1.0	TLS V1.2
1	17570	918162
2	52429	1576593
3	239701	1906271
4	105924	3789364
5	110334	6000882

Testing is done by doing the user access to several public services with several computers simultaneously to see the consistency of the system, using tools that have been prepared to monitor user activity.

Fig. 3 describes the test results referring to table 1, which is the result of testing encrypted data that is filtered based on the type of encryption used, namely tls v1.0 (black) and tls v1.2. (Red). In this study, data is displayed with several attributes, namely data_capture, total_lengt, windows, flags, protocols, and offsets that will be used in the clustering process.

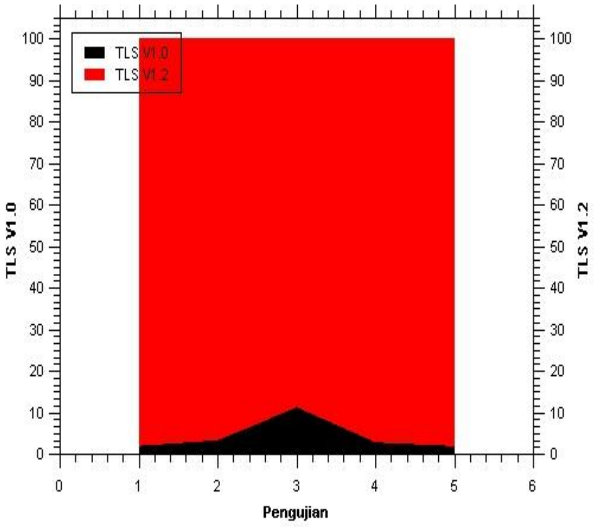

Fig. 3Encrypted Package Data Visualization

Table 2 presents cluster data for the TLSv1.2 data package types greater than TLS V1.0 data, this is because almost all public services use the TLS V1.2 protocol. on the TLS V1.0 cluster results in the third test, the data of 351 was obtained because many users access local web services. While in the trial for TLS V1.2 the most accessed is on the fifth test amounting to 5382631. Visualization of clusters of encrypted data packages can be seen in Fig. 4.

Table.2The results of the encrypted data packet cluster

Testing	Cluster	
	TLS V1.0	TLS V1.2
1	108	772198
2	126	1378485
3	351	1793955
4	162	3257128
5	141	5382631

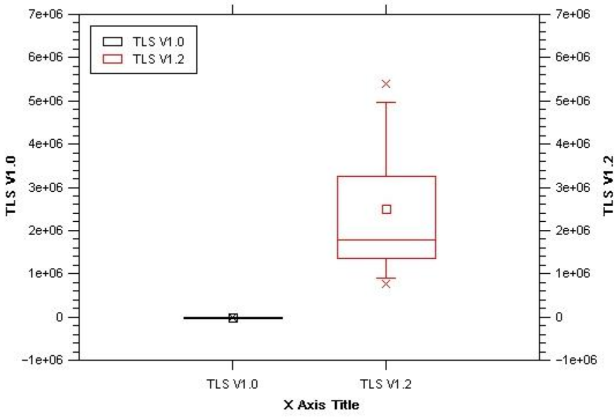

Fig. 4 Visualization of clusters of encrypted data packets

The cluster results are divided into two parts as shown in Fig. 4. The results of clustering data that have attribute patterns 127,221,111,237, AP and 16384 are TL1 package V1.0, while attribute patterns 127,112,98, 2046, AP and 0 are TL1 package V1.2.

The results of tests conducted are presented in Figure 4 shows that each cluster has a different detection level value, black is the shape of the cluster for testing data tls v1.0 and red color for tls v1.2. The results showed that the most data was tlsv1.2 because it rarely used the tls v1.0 application, even though the data obtained were presented in the form of numbers, but there were still errors in the introduction of data packet patterns.

After the clustering process is displayed in table 2 which is the total data cluster results from the encrypted clustering, then it will be calculated using a confusion matrix for the accuracy of each clustering result.

Table. 3Confusion Matrix

Binary Classification	Clustering result				
	1	2	3	4	5
TP	772198	1378485	1793955	3257128	5382631
FP	0	0	6	10	6
TN	145964	198108	112316	532236	618251
FN	0	0	0	0	0

V. CONCLUSION

The results of the visualization obtained show that the pattern of the packages is two versions of TLS, namely TLS 1.0 and TLS 1.2. The cluster method with BRICH produces significant data in grouping data packets. The cluster results result in more TLS V1.2 services than TLS v.1.0.

The accuracy obtained from the birch algorithm using a confusion matrix is 0.98 or 98 percent. This indicates that there are still 2% inaccuracies. Future studies will be developed with larger processing time and dataset and will be compared with other cluster algorithms so that they get better results.

REFERENCES

1. L. Deri, M. Martinelli, and A. Cardigliano, “nDPI : Open-Source High-Speed Deep Packet Inspection,” in International Wireless Communications and Mobile Computing Conference (IWCMC), 2014, pp. 617–622.
2. M. Husák, M. Čermák, T. Jirsík, and P. Čeleda, “HTTPS traffic analysis and client identification using passive SSL/TLS fingerprinting,” Eurasip J. Inf. Secur., vol. 2016, no. 1, pp. 1–14, 2016.
3. J. Lei, “An extended BIRCH-based clustering algoritm for large time-series datasets,” 2016.
4. S. Firdaus and M. Uddin, “A Survey on Clustering Algorithms and Complexity Analysis,” Int. J. Comput. Sci. Issues, vol. 12, no. 2, p. 62, 2015.
5. G. P. Spathoulas and S. K. Katsikas, “Enhancing IDS performance through comprehensive alert post-processing,” Comput. Secur., vol. 37, pp. 176–196, 2013.
6. A. Sanmorino, “A study for DDOS attack classification method,” J. Phys. Conf. Ser., vol. 1175, no. 1, 2019.
7. C. Meyer, “20 Years of SSL / TLS Research An Analysis of the Internet ’ s Security Foundation,” 2014.
8. S. Bhople, “Server based DoS vulnerabilities in SSL / TLS Protocols Master Thesis,” Eindhoven University of Technology, 2012.
9. H. Choi, H. Lee, and H. Kim, “Fast detection and visualization of network attacks on parallel coordinates,” Comput. Secur., vol. 28, no. 5, pp. 276–288, 2009.
10. R. Sánchez, Á. Herrero, and E. Corchado, “Visualization and clustering for SNMP intrusion detection,” Cybern. Syst., vol. 44, no. 6–7, pp. 505–532, 2013.