International Journal of Computer Networks and Communications Security

VOL. 3, NO. 3, MARCH 2015, 95–102
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

Media Storage Efficiency and Level Fingerprint Similarity in Network Forensic Analysis using Winnowing Multihashing Method

Irwan Sembiring[1], Jazi Eko Istiyanto[2], Edi Winarko[3] and Ahmad Ashari[4]

[1] Satya Wacana Christian University, Salatiga, Indonesia

[2, 3, 4] Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, GadjahMada, University, Yogyakarta, Indonesia

E-mail: [1]irwan@ staff.uksw.edu, [2]jazi@ugm.ac.id, [3]ewinarko@ugm.ac.id, [4]ashari@ugm.ac.id

ABSTRACT

Network forensics is a developing network security models that focused on the capture, recording, and analysis of network traffic, for the purposes of investigation. One of the problems in the Network forensics is the quantity or volume of data problems. Winnowing Multi hashing method can be used to conduct an investigation of attacks on the network forensic analysis. Value of Fingerprint is generated on Winnowing method Multi hashing (WMH), can be used as a marker of an attack that was captured by the Intrusion Detection System (IDS). WMH is a method that only takes excerpt of a payload. With this algorithm, the payload volume will be much more efficient because it only stores the fingerprint alone. This research is focused on the calculation of the efficiency of the storage medium and the optimum point combination fingerprint length, degree of similarity and storage media.

Keywords: *Winnowing Multi hashing, Jaccard Similarity, Network Forensic.*

1 INTRODUCTION

According to the agency Digital Forensics Research Workshop (DFRWS), digital forensic activities include preservation, collection, validation, identification, analysis, interpretation, documentation and presentation [1]. Because the equipment connected to the internet is increasingly a lot, then a forensic investigator will analyze the existing equipment, including Firewall, Intrusion Detection System (IDS), web server, and the real time traffic monitoring such as tcp dump or wire shark [2] [3]. There are five major problems on the complexity of digital forensic problems, problems of diversity, consistency and correlation, quantity or volume problems and Unified Time-lining problem [4]. An aspect of the volume (volume problem) becomes the focus in this research. Giura and Memon [5], the concluded research on capturing traffic on average 1300 flow in 1 second / second. Storage in a day it takes 10 GB, and 300 GB a month to reach 300 units by the number of hosts. On a scale WAN requires storage media as much as 1 TB / day. If this is maintained, certainly not efficient in terms of time and of storage media needs. The collection and storage of evidence in large volumes is a challenge. Many irrelevant data but still collected [6].Another way to analyze the payload is to find the unique pattern or feature extraction in a payload [7]. The pattern is then matched to obtain the degree of similarity. Winnowing Multi hashing method can be used to conduct an investigation of attacks on the network forensic analysis [8]. Fingerprint value generated on Winnowing method Multi hashing (WMH) can be used as a marker of an attack that was captured by the intrusion detection system (IDS). WMH is a method that only takes excerpt (excerpt) of a payload [9]. The main purpose of the method is to get the size of WMH more efficient storage

medium and fingerprint on the similarity percentage level alerts. Systematic in this paper include 1 Introduction, 2 Research Method, 3 Results and Anaysis and 4 Conclusion.

2 RESEARCH METHOD

Network forensics is a developing network security models that focused on the capture, recording, and analysis of network traffic, for the purposes of investigation [10]. Once the recording process is done, and then forwarded to the analysis. Generic network forensic models can be seen as Figure 1 [11].

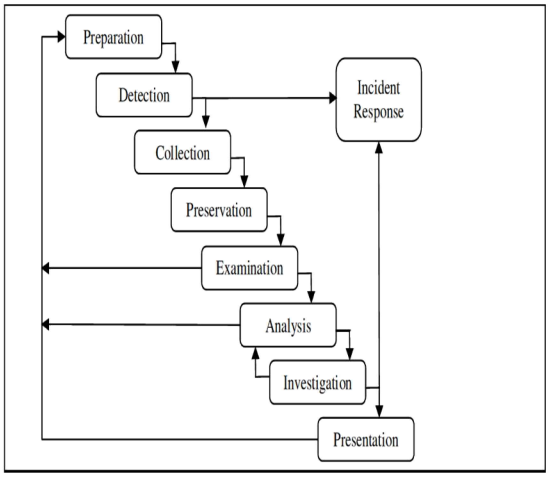

Fig. 1. Model Network forensics.

A. Preparation and Authorization.

Network forensic analysis focuses on network security devices such as Intrusion Detection System, Packet analysis, firewalls, and other support software. Network equipment security devices placed at strategic points of computer networks.

B. Detection of Incident / Crime.

Alert as a product network security tools that inform any abnormal traffic is a security breach or anomalies. Category and type of attack is determined based on certain parameters. Important validation of the alarm is false or not.

C. Incident Response

The response to crime or intrusion was detected beginning at the time the information is collected and validated. The response depends on the type of attack that is identified and organizational policies, laws and existing businesses.

D. Collection of Network Traces

Data obtained from the sensor used in capturing data traffic. The sensor should be safe, have limited access and should be able to avoid compromise. A standard procedure with reliable equipment, both hardware and software, should be placed to gather the maximum evidence.

E. Protection and Preservation

The original data were obtained in the form of traces and logs are stored in memory secondary. A hash of the data traces captured and protected. The standard procedure is used to ensure the accuracy and reliability of the data to perform preservation. Chain of custody must be maintained strictly so that no unauthorized use or tampering.

F. Examination

Analysis of reconstruction will be done thoroughly and integrated sensor data sources. Mapping and time lining needs to be done, so the most important data is not lost and does not mix. Data is hidden and camouflaged to be returned and classified in clustering in a group [12]. This mechanism facilitates the process of analysis in addition to also reduce the burden on storage media.

G. Analysis

Evidence has been collected and extracted. Indicator there is classified and correlated, to conclude an examination of patterns and types of attacks. Data mining and statistical approaches are often made reference to conduct this analysis. Some important parameters examined closely, such as fingerprint and DNS traffic. Attack patterns to be reconstructed simultaneously studied with a view to know who carried out the attack method.

H. Investigation and attribution

The information obtained from the trace evidence is used to identify who, what, where, when, how and why it happened. This will help in tracing back the source, the attack scenario reconstruction and attribution of sources. The most difficult part of network forensic analysis is to determine the identity of the attacker.

I. Presentation and review

The results will be presented with a good observation to be easily understood by managers of the organization. Explanation of all procedures used, displayed graphically, statistically, to support a conclusion.

In forensic analysis in accordance as in Figure 1, the method of trace back commonly used to find the attacker source . Analysis of the current trace back developed by considering the legitimacy, authenticity and integrity, such as trace back techniques Network Forensic Evidence Acquisition (NFEA) [13]. The trace back technique has two authentication schemes, known as Evidence Marking Scheme (AEMS) and Flow-based Selection Marking Scheme (FSMS) [13]. Winnowing algorithm is a derivative of the digital fingerprinting [14]. This algorithm was originally used for the benefit of copyright on the internet communication with XML. From the results of experiments conducted, which is the main characteristic of the winnowing is digital fingerprinting hide evenly on each partition. There are four basic properties are summarized in this study [14]:

1. Invisibility: By applying winnowing the data distortion will occur, however still provide useful and correct information to the user.
2. Preventing illegal embedding and verification: In winnowing algorithm, embedding and verification process is managed by a number of keys and data partition.
3. Blind verification: The original XML documents are not required in fingerprint verification.
4. Localization: By set up a fingerprinting, capable of detecting and narrow the modifications on a partition [12].

Winnowing with modifying the Rabin fingerprinting techniques [15], can detect all or most of the key documents. Each sequence of characters is stored in the storage array. Hashing is used to determine the marker as in Rabin fingerprinting. Winnowing Multi Hashing (WMH) is expected to reduce false positive circumstances existing on the query excerpt [9]. WMH not only provide good control on the size of the block, but also provide greater confidence to the query sequence excerpt on the overlap of existing blocks. Anomaly detection data packets can be seen from its payload. Most IDS to detect attacks on computer networks based on packet headers alone [14] proposed a data packet. .Anomaly focused on OSI layer. Payload is the actual data in the beyond data packet header. Header attached to the payload for transport purposes, and will be discarded after the package arrived at the destination. Payload data is collected and stored for offline processing. Actual historical data can be used for this purpose if it is available. From the entire data payload, the HTTP protocol is the threat of the highest candidate to be analyzed [16]. Winnowing algorithm is the basis of the reconstruction algorithm multi hashing. In internet crime, multi hashing winnowing algorithm used to extract the payload in the form excerpt called fingerprint. The purpose of this extraction is to measure the efficiency of the storage medium and the degree of similarity alerts. Experiments are conducted to extract useful information payload to detect an attack. This method has a better detection mechanism [17].

2.1 Winnowing Multihashing

Winnowing Multi Hashing method is one of the many methods that can be applied to this experiment of result payload efficiency show that this method produces a more significant improvement in the accuracy of quotations and data storage requirements compared to previous methods [16]. This method shows the best technique for selection on boundary block payload. At WMH method of determining the fingerprint using the following grammar:

1. Given a Hex, this is generally given as follows
$$S = s_1 s_2 s_3 \cdots s_n \ldots\ldots\ldots\ldots (1)$$
Where n is a lot of data.
2. The next step, to determine the size of k = k-gram, which is then used to form the Hash

$$T = (s_1 s_2 \cdots s_k), (s_2 s_3 \cdots s_{k+1}), \cdots, (s_{n-(k-1)} s_{n-(k-2)} \cdots s_{n-(k-k)}) \quad (2)$$

Looking for Hash value, by taking a prime number (p), then the calculation is given in Equation (3).

$$a_1 = (c_1 \cdot 16 + d_1)p^{k-1} + (c_2 \cdot 16 + d_2)p^{k-2} + \cdots + (c_k \cdot 16 + d_k)p^{k-k}$$

$$a_3 = (c_3 \cdot 16 + d_1)p^{k-1} + (c_2 \cdot 16 + d_2)p^{k-2} + \cdots + (c_{k+2} \cdot 16 + d_k)p^{k-k}$$

$$a_{n-(k-1)} = \left(c_{n-(k-1)} \cdot 16 + d_{n-(k-1)}\right)p^{k-1}$$

$$+(c_{n-(k-2)} \cdot 16 + d_{n-(k-2)}) \cdot p^{k-2} + \cdots + (c_{n-(k-k)} \cdot 16 + d_{n-(k-k)}) \cdot p^{k-k} \quad (3)$$

If taken r = n - (k -1), then the value obtained Hashing

$$A = \{a_1, a_2, \cdots, a_r\} \quad (4)$$

Each value in A, will be substituted on a function that is f (x) = xQ, where Q = (min A) - 1 then obtained

$$H = \{h_1, h_2, \cdots, h_r\} \quad (5)$$

Determining the value of a fingerprint based on Equation (5), which then each hash value will be the most sought smallest of any group Hash value (window size). Suppose a large window size is w.

$$U_1 = \{h_1, h_2, \cdots, h_w\}$$
$$U_2 = \{h_2, h_3, \cdots, h_{w+1}\}$$
$$U_3 = \{h_3, h_4, \cdots, h_{w+2}\}$$
$$\vdots$$
$$U_{r-(w-1)} = \{h_{r-(w-1)}, h_{r-(w-2)}, \cdots, h_{r-(w-w)}\} \quad (6)$$

Window size (N) will be in the partition as much r- (w-1) with r is much value Hash formed, and w is a lot of members in the U.

Determining the value of fingerprint

$$p_1 = \min U_1 = \min \{h_1, h_2, \cdots, h_w\}$$
$$p_2 = \min U_2 = \min \{h_2, h_3, \cdots, h_{w+1}\}$$
$$p_3 = \min U_2 = \min \{h_3, h_4, \cdots, h_{w+2}\}$$
$$\vdots$$
$$p_{r-(w-1)} = \min U_{r-(w-1)} = \min \{h_{r-(w-1)}, h_{r-(w-2)}, \cdots, h_{r-(w-w)}\} \quad (7)$$

Fingerprint is chosen is appropriate

$$K = \{y_i | Different\ Values\ of\ p_1 \cdots p_{r-(w-1)}\};\ i = 1 \cdots r - (w-1)$$

Hashing value on f_i taken as a fingerprint value, but need to be adjusted based on the hashing position starting from scratch. So that H in Equation , each sequence of numbers corresponding to H Hash will be

$$g_j = \{0, 1, 3, \cdots (n-1)\} \quad (8)$$

With regard to the order of different hash values (y_i) in Equation (8) and adjusted the order position based on Equation (9), it will obtain the sequence position numbers

$$Q = \{q_i | \ position\ of\ different\ value\ g_j\};\ 0 \le q_i \le n-1) \quad (9)$$

Furthermore, by using n-array relation (read: ener) consisting of 2 tuple (K, Q) represents the relationship between different hash value and position. So it can be written as

$$Fingerprint = F \subseteq K \times Q \quad (10)$$

or result in general will form

$$F = \{[y_i, q_i] | 0 \le q_i \le (n-1), 0 \le y_i \le (n-1)\} \quad (11)$$

A false positive occurs when querying against the elements x in hashing h1 ... hk applied to the value of x values obtained filtering is worth 1 If the hash value is assumed to be independent, then the probability to calculate the false positive rate (f) is as equation (13).

$$f = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k \approx \left(1 - e^{-\frac{kn}{m}}\right)^k \quad (12)$$

or can be reduced to the equation (13)

$$f = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k \approx \left(1 - \left(\lim_{x \to \infty} \left(1 - \frac{1}{x}\right)^{-x}\right)^{-\frac{kn}{m}}\right)^k \quad (13)$$

2.2 Detection of Similarity

Measurement of similarity fingerprint new payload degree as a new fingerprint with the fingerprint database that already exists. This percentage can be measured by the Jaccard Similarity Coefficient as in equation (15) [18]

$$D\ (A, B) = |\ A \cap B\ |\ /\ |\ A \cup B\ | \times 100 \quad (15)$$

Equation (15) describes the value of D (A, B) is the value of likeness or similarity, | A∩B | a pair of fingerprint intersection. | A∪B | is a number or a pair of fingerprint union. Similarities in the set S and T is the ratio between the slices and the union on the S and T so that it can be lowered by following equation (16)

$$Sim\ (C1, C2) = |\ S \cap T\ |\ /\ |\ S \cup T\ |\ x\ 100\% = Jaccard\ Similarity \quad (16)$$

For example if known c1 = {1, 2, 3} c2 = {1, 3, 4, 5} then the degree of similarity is = 40%.

2.3 Line of the Research

The line of this research is described as Figure 2 Payload captured by IDS in hexadecimal format will be extracted by the algorithm WMH. The output of WMH will generate a fingerprint mark as a keyword in a type of attack. Fingerprint stored for a false positive rate is calculated by considering k-grams and the window size to be determined by the user.

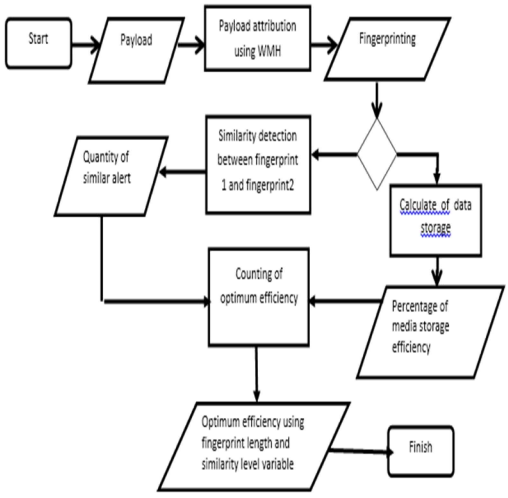

Fig. 2. Scheme Research

Each traffic of data captured by the IDS will have a fingerprint value. Each fingerprint will be matched with others fingerprint to measure the degree of similarity. Similarity values used in this study using Jaccard Similarity techniques. The process of matching a fingerprint on the alert with all alerts is limited to one type of attack classification. The attack Classification type was captured by IDS as shown in Figure 3.

Classification	Total #	Sensor #	Signature	Source Address	Dest. Address	First	Last
unclassified	2162 (1%)	1	3	106	5	2013-11-06 07:14:43	2013-12-02 15:35:49
attempted-dos	8726 (4%)	1	11	345	25	2013-05-18 15:31:21	2014-08-14 10:49:53
attempted-user	27725 (13%)	1	52	729	6	2013-05-19 04:01:37	2014-03-12 07:55:11
attempted-recon	3240 (2%)	1	10	100	13	2013-06-22 17:09:38	2014-08-14 12:22:03
misc-activity	7422 (4%)	1	14	298	8	2013-06-24 15:33:12	2014-08-14 08:48:23
policy-violation	3651 (2%)	1	7	653	2	2013-06-03 13:25:51	2014-03-12 07:33:02
misc-attack	315 (0%)	1	4	102	5	2013-06-04 17:15:12	2014-07-10 14:57:23
web-application-attack	691 (0%)	1	1	1	8	2013-06-06 12:17:58	2013-10-08 14:24:27
trojan-activity	17 (0%)	1	4	4	4	2013-06-06 16:54:01	2014-04-06 15:06:41
attempted-admin	173 (0%)	1	5	38	7	2013-06-10 08:49:23	2014-08-14 04:31:08
bad-unknown	15 (0%)	1	1	2	2	2013-10-01 10:24:45	2014-02-05 23:30:13
successful-admin	1 (0%)	1	1	1	1	2013-10-08 13:37:11	2013-10-08 13:37:11
successful-user	1 (0%)	1	1	1	1	2013-10-08 11:19:57	2013-10-08 11:19:57
icmp-event	145805 (70%)	1	1	33	6	2013-11-07 06:25:28	2014-08-13 06:47:31
shellcode-detect	9161 (4%)	1	7	577	2	2013-11-08 07:02:11	2014-07-31 11:40:26
system-call-detect	137 (0%)	1	2	76	2	2013-11-08 10:04:08	2014-03-12 07:27:28
network-scan	299 (0%)	1	1	48	9	2013-11-16 05:46:42	2014-08-14 11:14:19

Fig. 3. Classification of Attacks

In Figure 3, all the traffic captured by the IDS within a period of 1 year from a total of 209 341 alerts are categorized into 17 types include DOS attacks, Trojans, Web Attack, ICMP Attack, Scanning and others. If not detected as an attack that has been registered in the database, the IDS will be categorized as unclassified.

3 RESULTS AND ANALYSIS

3.1 Measurement of the storage media efficiency

The basic idea of this measurement is how much efficiency is gained using fingerprint, as a representation of the method of WMH. In accordance with the framework of network forensic preservation and collection stages as identified in Figure 1 Data derived from IP address 124.81.113.178, this data has been validated with a checksum. Additional information supporting the attack time is 11:29:07 on the 25th November 2013.

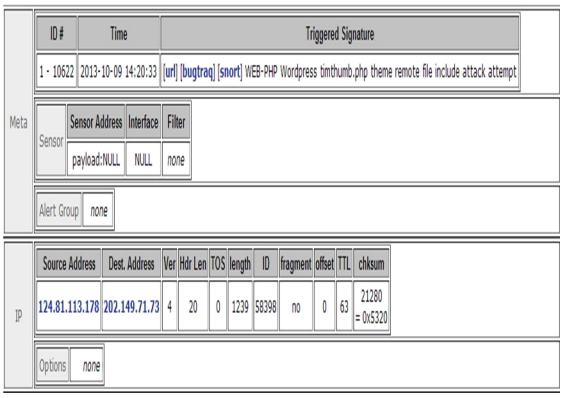

Fig. 4. Metadata Web Attacks

Examples of Web attacks is shown in Figure 4. Bootnet is in the category of Trojan attacks. The volume of data in a single alert is 1187 Bytes. The format of the data captured in the form of a hexadecimal representation of the data link layer, shown in Figure 5.

```
47 45 54 20 2f 66 61 62  6c 65 2f 66 61 62 6c 65
2d 66 6f 61 2d 32 2e 36  32 37 33 2e 7a 69 70 20
48 54 54 50 2f 31 2e 31  0d 0a 41 63 63 65 70 74
3a 20 2a 2f 2a 0d 0a 41  63 63 65 70 74 3a 20 2a
2f 2a 0d 0a 52 61 6e 67  65 3a 20 62 79 74 65 73
3d 30 2d 31 31 31 39 0d  0a 55 73 65 72 2d 41 67
65 6e 74 3a 20 4d 6f 7a  69 6c 6c 61 2f 34 2e 30
20 28 63 6f 6d 70 61 74  69 62 6c 65 3b 20 29 0d
0a 48 6f 73 74 3a 20 72  65 73 2e 63 6f 74 2e 79
65 65 70 67 61 6d 65 2e  63 6f 6d 0d 0a 43 6f 6f
6b 69 65 3a 20 5f 5f 75  74 6d 61 3d 39 33 37 35
37 35 30 36 2e 31 37 32  30 34 34 30 39 33 36 2e
31 33 38 33 36 36 34 39  33 36 2e 31 33 38 33 36
36 34 39 33 36 2e 31 33  38 33 36 36 34 39 33 36
2e 31 3b 20 5f 5f 75 74  6d 7a 3d 39 33 37 35 37
35 30 36 2e 31 33 38 33  36 36 34 39 33 36 2e 31
2e 31 2e 75 74 6d 63 73  72 3d 28 64 69 72 65 63
74 29 7c 75 74 6d 63 63  6e 3d 28 64 69 72 65 63
74 29 7c 75 74 6d 63 6d  64 3d 28 6e 6f 6e 65 29
0d 0a 58 2d 50 72 6f 78  79 2d 49 44 3a 20 31 38
36 32 39 36 30 36 34 33  0d 0a 58 2d 46 6f 72 77
61 72 64 65 64 2d 46 6f  72 3a 20 31 39 32 2e 31
36 38 2e 37 2e 31 37 39  0d 0a 56 69 61 3a 20 31
2e 31 20 31 39 32 2e 31  36 38 2e 37 2e 32 35 34
20 28 4d 69 6b 72 6f 74  69 6b 20 48 74 74 70 50
72 6f 78 79 29 0d 0a 0d  0a
```

Fig. 5. Footage Payload of Web attacks

The Results of extraction with WMH produce false positive rate is the maximum combined 0:01, with k g = 6, the window size is set to a value trend graph 128 false positive rate with the combination of k-gram values can be seen in Figure 6 payload capacity of 1187 bytes.

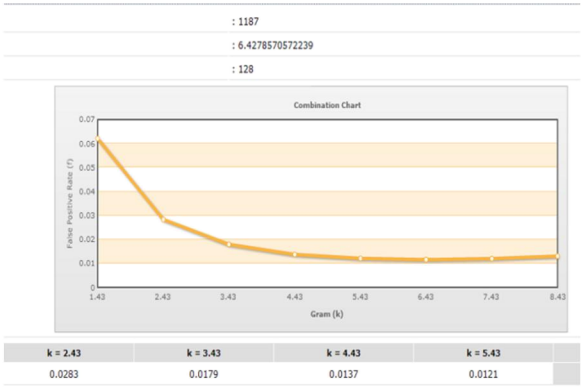

Fig. 6. Combination of False Positive Rate Fingerprint

Results winnowing algorithm that will generate the fingerprint is

[1,119**][**104553,245**][**116565,280**][**151550,298**][**172332,300**][**494933,357**][**831935,427**][**78622,516**][**12965,586**][**255330,587**][**710899,630**][**72943,734**][**124637,748**][**125704,844**][**39141,911**][**270559,969**][**494385,1062**][**285890,1103**][**39372,1152**][**542490,1153**][**6077940,1154**][**15742365,1156**][**15932474,1158**]**

The fingerprint in hexadecimal format is:

30: 20: 30: 20: 70: 20: 70: 20: 30: 30: 20: 30: 30: 30: 30: 30: 20: 70: 50: 72: 78: 29

One fingerprint block is formed consisting of a fingerprint value and offset value. From 23 bytes extracted fingerprint, If the big unknown payload capacity of the payload (payload length) is (P), for example 1187 bytes. Unique alert (U) is 125, the length of the fingerprint (F) is 23, while the total alerts (A) The amount of the Storage media efficiency percentage reaches 209712 as the equation (17).

$$E = \frac{(Px\ A)-\left((P+Fx\ U)+(F\ x\ A)\right)}{(Px\ A)}\ x\ 10 \qquad (17)$$

$$= \frac{(1187x\ 209712)-(\ 1187+23)x\ 125+23\ x\ 209712}{1187\ x\ 209712}\ x\ 100$$

$$= \frac{(1187x\ 6\ 6)-(\ 1187+23)x\ 125)+(23\ x\ 6\ 6\ 1}{1187\ x\ 6\ 6\ 1}\ x\ 100$$

= 243953518 /248928144 = 0.98 =**98 %**

= 768194 /784607 = 0.98 =**98 %**

Equation (17), illustrates the level of efficiency obtained by using WMH method. Comparison between unique alerts, fingerprint and total alerts successfully captured by IDS is the basis of high and low levels of efficiencies gained. Value of 98% obtained from the magnitude of the difference between the number of alerts that compared with the 209 712 unique alerts (signature) of = 125 Trial Results as Figure 6 describes the combination of total alerts (A) / number of unique alerts (U) .From experimental and simulation results performed if the total alerts (a) ≥4% of the amount of unique alerts, it will acquire a positive value efficiency trends.

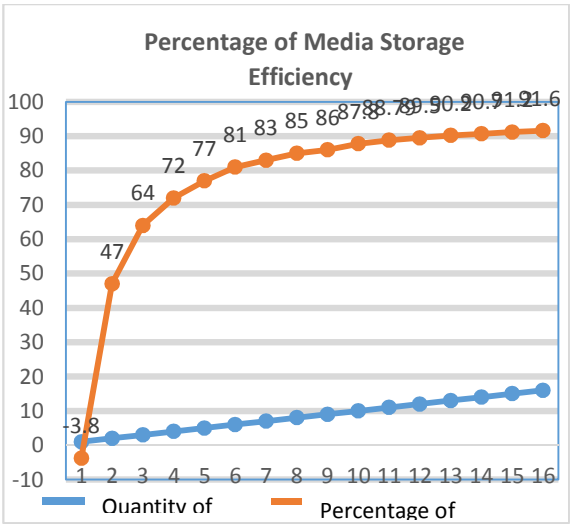

Fig. 7. Trends in Media Storage Efficiency

Figure 7, describes the trend of the efficiency magnitude obtained with the ratio between the total alerts and unique alerts. For example if the total alert is five times larger than the unique alerts, efficiency obtained is 77%. From the results of experiments carried out, if the fingerprint length = 23 bytes, then the efficiency will occur if the total alerts least 4% of unique alerts. Calculation of storage efficiency on type attack classification also produces the same percentage. In this experiment contained a total of 661 attacks in web attacks category, with unique alerts = 1 If the fingerprint length is 23 bytes and the length of the alerts that are detected is 1187 bytes, the efficiency of the obtained is 98%.

From equation (17), if the condition of the payload length is not the same then the equation can be derived as equation (18).

$$\frac{\sum payload - (\sum Tabel\ Payload + \sum Fingerprint)}{\sum payload}\ x\ 10 \quad (18)$$

3.2 Percentage of Similarity

Measuring the similarity percentage level is an important part in this research. In experiments initotal web alerts attack 661 times. If the alerts such as alerts numbers 10382 compared to 660 alerts the others, then found the amount of alerts that have a similarity score> = 80 percent is as much as 68 alerts.

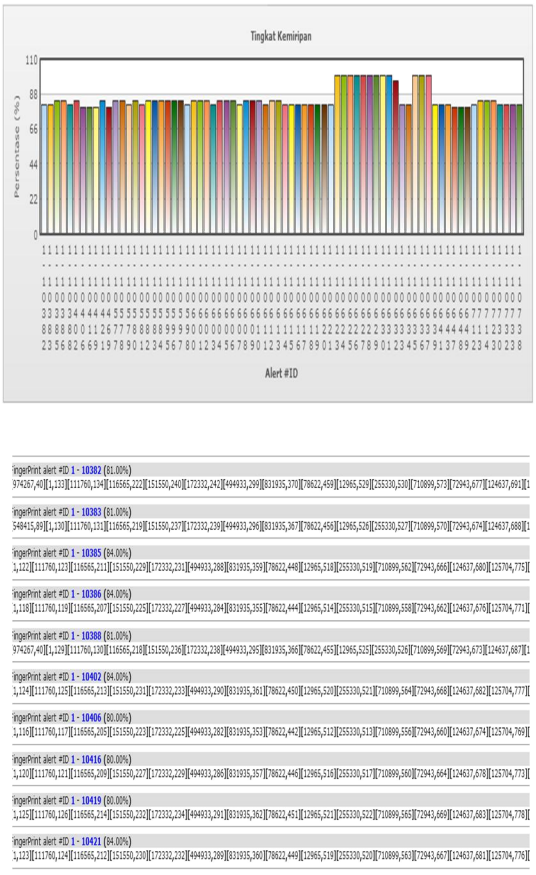

Fig. 8. Percentage of similarity

3.3 Relationship between Similarity, Fingerprint and Efficiency

The results of experiments conducted with the implementation of fingerprint IDS obtained a high enough efficiency to reach 98%. Windows size = 128 then, fingerprint length = 23, 80% above the level of similarity with a total of 661 amounted to 80 alerts Relations storage efficiency with the combination of a shift in the level of similarity, the value of the fingerprint on the payload length of 1187 bytes as shown in Table 1.

Table 1: Relationship between Similarity, Fingerprint and Efficiency in Web Attack

Window size	Fingerprint	False Positive	Efficiency (%)	Similarity >80 %
128	23	0.016	98	68
100	25	0.005	97,8	68
80	30	0,002	97,4	72
64	38	0.0007	96,7	72

In Table 1, the variations in windows size are tested randomly. This value will affect the length of the fingerprint. With value k-gram = 6 then the combination of the efficiency and value of similarity can be seen as Table 1 In Table 2 the results of an experiment to search for the inherent similarity and minimal alerts efficiency .Total (A) ≥4% of unique alerts. If the alert is unique (U) is obtained from this experiment is 125, then A = 130 intersection point between similarity and efficiency as Table 2 and Figure 8.

Table 2: Relationship Similarity, Fingerprint and Efficiency in (A) ≥4%, N = 125

Window size	Fingerprint	False Positive	Efficiency	Similarity >80 %
128	23	0.016	27 %	10 %
100	25	0.005	26 %	10 %
80	30	0,002	26 %	10 %
64	38	0.0007	25 %	11 %
32	75	0.00005	19%	11%
10	197	0.00001	3%	11%

Fig. 8. Inherent between the magnitudes of the similarity efficiency

In Figure 8, the intersection graphs of Table 2 shows intersection between the efficiency level and the similarity percentage from the inherent value = 0.11. This value is the condition of the payload

length = 1187 bytes, fingerprint 197 with window size = 10.

4 CONCLUSION

1. By using the WMH method, then began optimal storage efficiency if the number of alerts minimum of> 4% of the total signatures.
2. The smaller the size of windows, the greater the value of the fingerprint and will certainly get a false positive rate getting smaller. High value of the fingerprint effect on the lower level of efficiency, but the level of similarity will be higher.
3. By using the WMH method, the average efficiency obtained in this experiment was 97%. This value depends on the gap between high and low total alerts to total signatures.

7 REFERENCES

[1] DFRWS Technical Committee. (DFRWS)., "A Road map for Digital Forensic Research : DFRWS Technical Report ", DTR - T001-01 FINAL , 2004

[2] V. Corey, et al., "Network Forensics analysis ", IEEE Internet Computing Institute of Electrical and Electronic Engineers IEE Explore, page 60, November-December 2002.

[3] S. Mukkamala, A and H. Sung., "Identifying significant Features for Network Forensic Analysis Using Artificial Intelligent Techniques", International Journal of Digital Evidence, Vol. 1, Issue 4, Winter 2003.

[4] Raghavan, S.," Digital Forensic Research: Current State-of-the-Art ", Springer CSIT, 1 (1): 91–114, March 2013

[5] Giura, P., & Memon, N., "Efficient Methods to Store and Query Network Flow Data" Polytechnic Institute of NYU, Department of Computer Science and Engineering, New York, 2011.

[6] Kaushik, K.A, et al., Network Forensic System for Port Scanning Attack, @ IEEE, 2010

[7] Hao, F, et al., "Fast Payload-Based Flow Estimation for Traffic Monitoring and Network Security", Copyright ACM, 2013

[8] Sembiring, I et al., "Payload Attribution Using Winnowing Multi Hashing Method,",International Journal of Information & Network Security (IJINS) ,Vol.2, No.5, October 2013, pp. 360~370 , ISSN: 2089-3299, 2013.

[9] M.Ponec, et al., "New Payload Attribution Methods for Network Forensic Investigations," ACM Transactions on Information and System Security, Vol. 13, No. 2, Article 15, Publication, February 2010.

[10] Almulhem, A.and Issa, T., " Experience with Engineering a Network Forensics System" ISOT Research Lab University of Victoria, Canada., 2004.

[11] Yusoff, Y, et al.,"Common Phase of Computer Forensic Investigation Model" , International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 3, 2004.

[12] Beverly, R., et al.,"Forensic carving of network packets and associated data structures" , Naval Postgraduate School Monterey California, United States, 2011.

[13] Kim, S.H and Kim, K.H., Network Forensic Evidence Acquisition (NFEA) With Packet Marking, © IEEE, 2011

[14] Darwish, S.M.,"New system to fingerprint extensible markup language documents using winnowing theory" IET Signal Process., Vol. 6, Iss. 4, pp. 348 – 357, 2012.

[15] Schleimer, S., et al., "Winnowing Local Algorithms for Document Fingerprinting" ,Proceedings of the 2003 ACM, SIGMOD International Conference on Management of Data (SIGMOD'03), ACM, New York, 76–85, 2003.

[16] Hongcheng, T., and Jun Bi., "An Incrementally Deployable Flow-Based Scheme for IP Traceback", IEEE Communications Letters 16, 1140-1143, 2012.

[17] Mrdovic,S, et al.," Combining Static and Live Digital Forensic Analysis in Virtual Environment , IEEE, 2009.

[18] Bank, J and Cole, B., " Calculating the Jaccard Similarity Coe_cient with Map Reduce for Entity Pairs in Wikipedia" , Wikipedia Similarity Team 2013.