International Journal of Information & Network Security (IJINS)
Vol.3, No.6, December 2014, pp. 359~369
ISSN: 2089-3299

The Use of Winnowing Multihashing Method for the Media Capacity Efficiency in Network Forensic Analysis

Irwan Sembiring, Jazi Eko Istiyanto, Edi Winarko, Ahmad Ashari
Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Gadjah Mada University, Yogyakarta, Indonesia

Article Info

Article history:

Received Jun 12th, 2014
Revised Aug 20th, 2014
Accepted Aug 26th, 2014

Keyword:

Winnowing Multi hashing
Jaccard Similarity
Network Forensic.

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.

Corresponding Author:

Irwan Sembiring,
Faculty of Information Technology, Satya Wacana Christian University, Diponegoro 52-60 Salatiga – Jawa Tengah Indonesia. Email : irwan@staff.uksw.edu or irwan.sembiring@ugm.ac.id

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, then forwarded to the analysis. Generic network forensic models can be seen as Figure 1 [11].

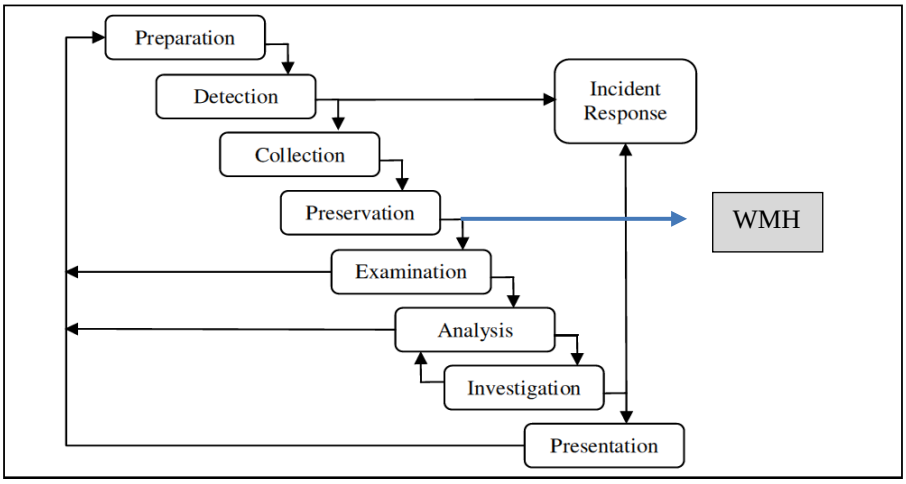


Figure 1 Model Network forensics.

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.

According to Figure 1, forensic analysis using the trace back method is commonly used to find the attacker . Network Forensic Evidence Acquisition (NFEA) is the method developed from traceback method by considering the legitimacy, authenticity and integrity[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. Results of the conducted experiments, showed that the main characteristisc of the winnowing is hidden repetitive fingerprinting on each block boundary. There are four basic properties 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.

Fingerprinting located on the paylod can detect the modifications or authenticity of a partition precisely[12].

Winnowing, derived from Rabin Fingerprinting modification[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]. Not only provides good control on the size of the block, WMH also provides query sequence excerpt on the overlapping blocks accurately. Anomaly detection can be carried on its payload. The IDS commonly used to detect the intrusion is based only on packet headers[14] . The next level of IDS should detect the intrusion on each 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 traffic, HTTP protocol service is the of the most attacked protocol [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 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. Fingerprint determining on WMH are conducted as follows:

1. A hexadecimals is generally given as follows

$$S = s_1 s_2 s_3 \cdots s_n \quad \text{.....(1)}$$

n is number of data.

2. The next step is to determine the size of k = k-gram, which will be used to form the Hash, according to equation (2)

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

3 To determine the Hash value, take a prime number (p), then calculate the number using Equation (3).

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

$$a_2 = (c_2 \cdot 16 + d_1)p^{k-1} + (c_3 \cdot 16 + d_2)p^{k-2} + \cdots + (c_{k+1} \cdot 16 + d_k)p^{k-k}$$

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

$$a_{n-(k-1)} = (c_{n-(k-1)} \cdot 16 + d_{n-(k-1)})p^{k-1} + (c_{n-(k-2)} \cdot 16 + d_{n-(k-2)}) \cdot p^{k-2} + \cdots + (c_{n-(k-k)} \cdot 16 + d_{n-(k-k)}) \cdot p^{k-k} \quad (3)$$

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.

$$U_{r-(w-1)} = \{h_{r-(w-1)}, h_{r-(w-2)}, \cdots, h_{r-(w-w)}\} \quad (6)$$

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

$$K = \{y_i | Different\ Values\ of\ p_1 \cdots p_{r-(w-1)}\};\ i = 1 \cdots r - (w-1) \quad (8)$$

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 (8), each sequence of numbers corresponding to H Hash will be

$$g_j = \{0, 1, 3, \cdots (n-1)\} \quad (9)$$

With regard to the order of different hash values (y_i) in Equation (9) and adjusted the order position based on Equation (3.10), it will obtain the sequence position numbers

$$Q = \{q_i | \ position\ of\ different\ value\ g_j\};\ 0 \le q_i \le n-1) \quad (10)$$

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

$$Fingerprint = F \subseteq K \times Q \quad (11)$$

or result in general will form

$$F = \{[y_i, q_i] | 0 \le q_i \le (n-1), 0 \le y_i \le (n-1)\} \quad (12)$$

A false positive used in WMH 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) .

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

or can be reduced to the equation 3:14

$$f = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k \approx \left(1 - \left(\lim_{x \to \infty} \left(1 - \frac{1}{x}\right)^{-x}\right)^{-\frac{kn}{m}}\right)^k \quad (14)$$

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].

$$Z(X, Y) = | X \cap Y | / | X \cup Y | \times 100 \quad (15)$$

Equation (15) describes the value of Z (X, Y) is the value of similarity,X is the fingerprint on first payload, Y is the fingerprint on second payload. | X∩Y |is a pair of fingerprint intersection. | X∪Y | 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)

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

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

2.3 SCHEME OF 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.

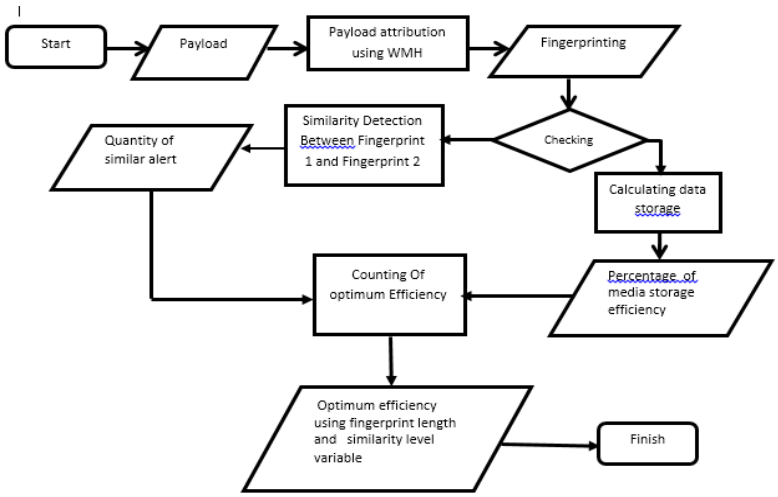


Figure 2 Scheme of Research

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 is based on 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	109	5	2013-11-06 07:14:49	2013-12-02 15:36:49
attempted-dos	8726 (4%)	1	11	345	35	2013-05-18 15:31:21	2014-08-14 10:49:53
attempted-user	27735 (13%)	1	52	739	6	2013-05-19 04:01:37	2014-03-12 07:55:11
attempted-recon	3242 (2%)	1	10	100	13	2013-08-22 17:08:38	2014-08-14 12:22:03
misc-activity	7423 (4%)	1	14	390	8	2013-08-24 15:39:12	2014-08-14 08:48:23
policy-violation	3651 (2%)	1	7	653	2	2013-09-03 13:26:51	2014-03-12 07:38:02
misc-attack	315 (0%)	1	4	102	5	2013-09-04 17:15:12	2014-07-18 14:57:23
web-application-attack	661 (0%)	1	1	1	8	2013-09-06 13:17:58	2013-10-09 14:24:27
trojan-activity	17 (0%)	1	4	4	4	2013-09-09 16:54:01	2014-01-09 15:06:41
attempted-admin	179 (0%)	1	5	38	7	2013-09-10 08:49:23	2014-08-14 04:31:09
bad-unknown	15 (0%)	1	1	2	2	2013-10-01 18:24:45	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	2013-10-09 11:10:57	2013-10-09 11:10:57
icmp-event	145695 (70%)	1	1	33	6	2013-11-07 06:25:28	2014-08-13 08:47:31
shellcode-detect	9161 (4%)	1	7	577	2	2013-11-09 07:02:11	2014-07-31 11:40:26
system-call-detect	137 (0%)	1	2	76	2	2013-11-09 19:04:09	2014-03-12 07:27:20
network-scan	219 (0%)	1	1	48	9	2013-11-16 05:46:42	2014-08-14 11:14:19

Figure 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 such as DOS attacks, Trojans, Web Attack, ICMP Attack, Scanning and among others. If not detected as an attack that has been registered in the database,then the output of 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.

Meta

ID #	Time	Triggered Signature
1 - 10622	2013-10-09 14:20:33	[url] [bugtraq] [snort] WEB-PHP Wordpress timthumb.php theme remote file include attack attempt

Sensor	Sensor Address	Interface	Filter
	payload:NULL	NULL	*none*

Alert Group	*none*

IP

Source Address	Dest. Address	Ver	Hdr Len	TOS	length	ID	fragment	offset	TTL	chksum
124.81.113.178	202.149.71.73	4	20	0	1239	58398	no	0	63	21280 = 0x5320

Options	*none*

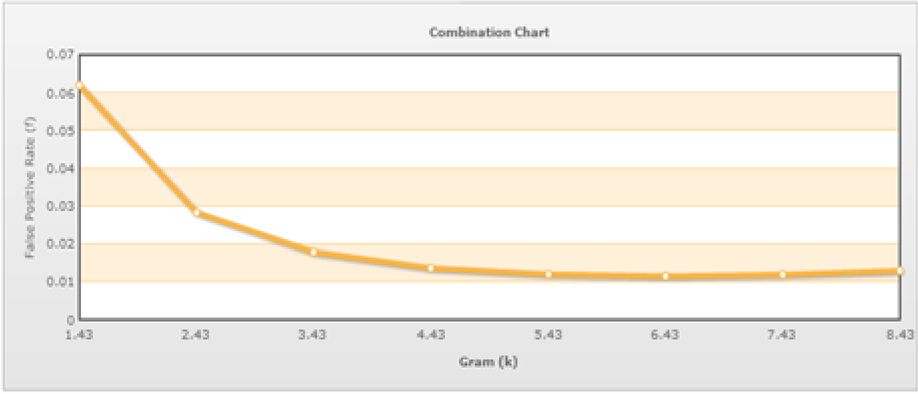
Figure 4 Metadata Web Attacks

Examples of Web attacks is shown in Figure 4. Botnet 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 is 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
```

Figure 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.



k = 2.43	k = 3.43	k = 4.43	k = 5.43	k = 6.43
0.0283	0.0179	0.0137	0.0121	0.0116

Figure 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: 30: 20: 70: 50: 72: 78: 29
False positive rate value determined from WMH is 0.0116, as seen on Figure 6

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).

$$E = \frac{(P x\ A) - ((P + F x\ U) + (F\ x\ A))}{(P x\ A)}\ x\ 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (17)$$

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

$$= \ 243953518\ /248928144 = 0.98\ \ = \mathbf{98\ \%}$$

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.

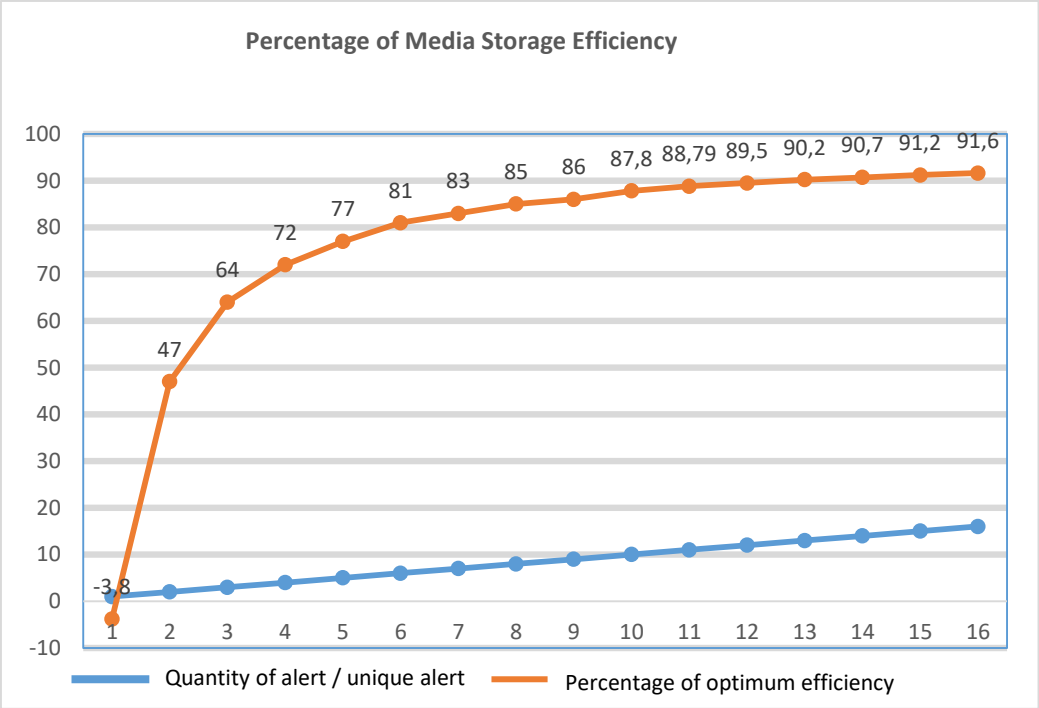


Figure 7. Trends in Media Storage Efficiency

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%.

$$= \frac{(1187x\ 660)-(\ 1187+23)x\ 125)+(23\ x\ 661\ )}{1187\ x\ 661}\ x\ 100$$

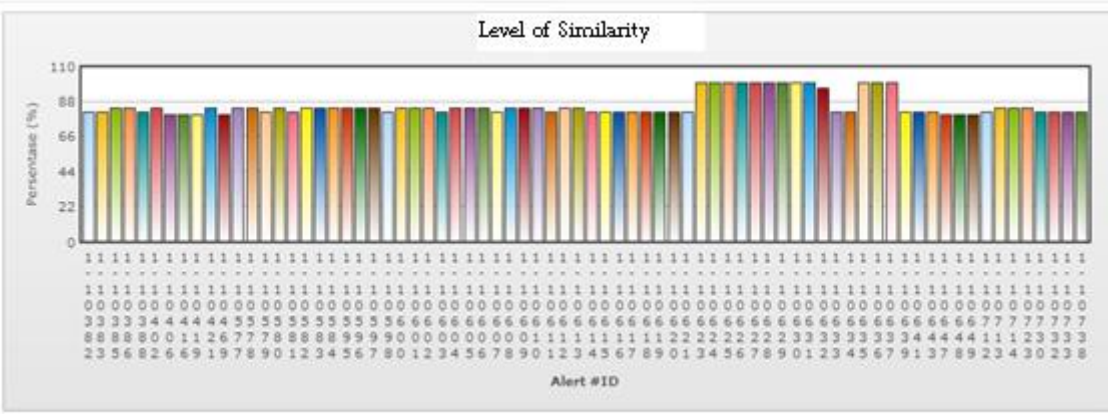
$$= \quad 768194\ /784607 = 0.98 \ = \mathbf{98\ \%}$$

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

$$\frac{\sum payload-(\sum Tabel\ Payload+\sum Fingerprint)}{\sum payload}\ x\ 100\text{……..(18)}$$

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.



FingerPrint alert #ID **1 - 10382** (81.00%)
:974267,40][1,133][111760,134][116565,222][151550,240][172332,242][494933,299][831935,370][78622,459][12965,529][255330,530][710899,573][72943,677][124637,691][1

FingerPrint alert #ID **1 - 10383** (81.00%)
:548415,89][1,130][111760,131][116565,219][151550,237][172332,239][494933,296][831935,367][78622,456][12965,526][255330,527][710899,570][72943,674][124637,688][1

FingerPrint alert #ID **1 - 10385** (84.00%)
:1,122][111760,123][116565,211][151550,229][172332,231][494933,288][831935,359][78622,448][12965,518][255330,519][710899,562][72943,666][124637,680][125704,775][

FingerPrint alert #ID **1 - 10386** (84.00%)
:1,118][111760,119][116565,207][151550,225][172332,227][494933,284][831935,355][78622,444][12965,514][255330,515][710899,558][72943,662][124637,676][125704,771][

FingerPrint alert #ID **1 - 10388** (81.00%)
:974267,40][1,129][111760,130][116565,218][151550,236][172332,238][494933,295][831935,366][78622,455][12965,525][255330,526][710899,569][72943,673][124637,687][1

FingerPrint alert #ID **1 - 10402** (84.00%)
:1,124][111760,125][116565,213][151550,231][172332,233][494933,290][831935,361][78622,450][12965,520][255330,521][710899,564][72943,668][124637,682][125704,777][

FingerPrint alert #ID **1 - 10406** (80.00%)
:1,116][111760,117][116565,205][151550,223][172332,225][494933,282][831935,353][78622,442][12965,512][255330,513][710899,556][72943,660][124637,674][125704,769][

FingerPrint alert #ID **1 - 10416** (80.00%)
:1,120][111760,121][116565,209][151550,227][172332,229][494933,286][831935,357][78622,446][12965,516][255330,517][710899,560][72943,664][124637,678][125704,773][

FingerPrint alert #ID **1 - 10419** (80.00%)
:1,125][111760,126][116565,214][151550,232][172332,234][494933,291][831935,362][78622,451][12965,521][255330,522][710899,565][72943,669][124637,683][125704,778][

FingerPrint alert #ID **1 - 10421** (84.00%)
:1,123][111760,124][116565,212][151550,230][172332,232][494933,289][831935,360][78622,449][12965,519][255330,520][710899,563][72943,667][124637,681][125704,776][

Figure 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) more than 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%

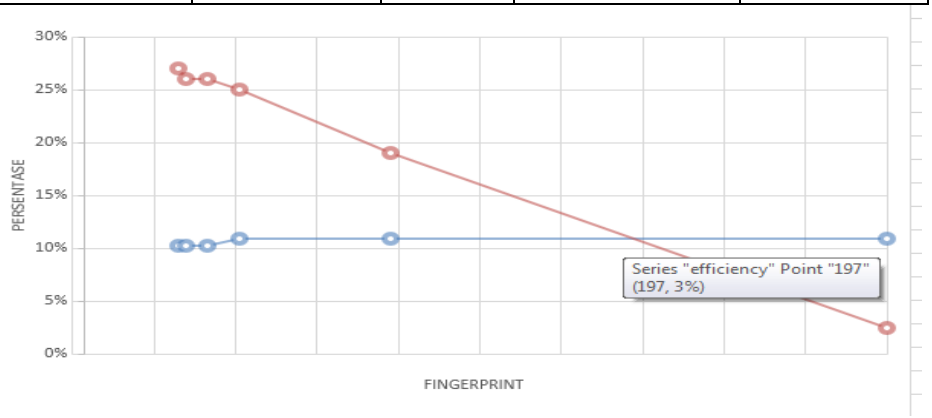


Figure 8. Inherent between the magnitudes of the similarity efficiency

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, the optimal storage efficiency will have a positive values if the number of alerts is more than 4% of the total signatures.
2. The smaller the windows size, the greater the value of the fingerprint. The greater the fingerprint, the smaller the false positive rate. 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 storage efficiency obtained in this experiment was 97%. This value depends on the gap between high and low total alerts to total signatures.

REFERENCE

[1] DFRWS Technical Committee. (DFRWS)., "A Road map for Digital Forensic Research : DFRWS Technical Report ", *DTR - T001-01 FINAL* , 2004
[2] Corey,V.,Peterman, C., Shearin, S., Greenberg, M.S., Bokkelen, J.,V., "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, A.K.,Pilli,E.S., & Joshi, R.C, Network Forensic System for Port Scanning Attack, *IEEE*, 2010
[7] Fang Hao, Murali S. Kodialam, T. V. Lakshman, Hui Zhan "Fast Payload-Based Flow Estimation for Traffic Monitoring and Network Security", *ACM*, 2013

[8] Sembiring, I.,Istiyanto, J.E., Ashari, A., Winarko, E., "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] Ponec, M., Giura, P., Wein, J., & Onnimann, H., "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., Ismail, R. and Hassan, Z.,"Common Phase of Computer Forensic Investigation Model" , *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol 3, No 3, 2004.
[12] Beverly, R., Simson, G. and Greg, C.,,"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., Wilkerson D., &S.,Aiken, A., "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., Huseinovic,A., Zajko, E., 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.

BIBLIOGRAPHY OF AUTHORS

Irwan Sembiring, completed his undergraduate program in UPN " Veteran" Yogyakarta, majoring inInformation Technology in 2001, pursued higher degree in School of Computer Science and Electronics Gadjah Mada University ,Yogyakarta, Indonesia and received Master Computer in 2004.Now he is a doctoral candidate in School of Computer Science and Electronics Gadjah Mada University ,Yogyakarta, Indonesia . Hisresearch interests include Network Security and Digital Forensic. Email irwan@staff.uksw.edu and irwan.sembiring@ugm.ac.id.

Jazi Eko Istiyanto received a B.Sc in **Physics** (1986) from Gadjah Mada University, Yogyakarta, Indonesia. He then pursued higher degrees in the the University of Essex, UK and received a **Postgraduate Diploma in Computer Programming and Microprocessor Applications**(1987), an **M.Sc in** Computer Science(1988), and a **Ph.D in** Electronic Systems Engineering (1995), all from the University of Essex. His research interest covers information security, electronic systems optimization, and embedded systems.he is a Professor of Electronics and Instrumentation. Email jazi@ugm.ac.id

Edi Winarko, lecturer at the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjahmada University. I received my S1 degree in Statistics from Gadjahmada University, MSc. in Computer Sciences from Queen's University, Canada, and Ph.D in Computer Sciences from Flinders University, Australia. His research interest covers Data Warehousing and Data Mining Information Retrieval. Email ewinarko@ugm.ac.id

Ahmad Ashari, received a B.Sc in Physics (1988) from Gadjah Mada University, Yogyakarta, Indonesia. He then pursued higher degree and received a Master Computer (1992)from Universitas Indonesia, Jakarta, Indonesia and Ph.D in Informatics engineering (2001) from Vienna University of Technology Austria. His research interest covers data communication and computer network , internet and *www*, and distributed and parallel computing systems. Email ashari@ugm.ac.id