International Journal of Information & Network Security (IJINS)
Vol.2, No.5, October 2013, pp. 360~370
ISSN: 2089-3299

Payload Attribution Using Winnowing Multi Hashing Method

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, 201x
Revised Aug 20th, 201x
Accepted Aug 26th, 201x

Keyword:

Payload
Winnowing Multi Hashing
False Positive

ABSTRACT

Payload attribution is a process to identify the sources and destinations of all packets that appeared on a network and a certain excerpt of a payload. This method can be used for traffic efficiencies in investigating internet crime (cybercrime), such as tracing who is responsible for activities for unauthorized access, illegal contents, deliberate spread of the virus, data forgery and any cybercrime. The payload is the actual data that is sent by the packet to the destination. The aim using Winnowing Multi Hashing Method (WMH) is to extract the payload by calculating the value of false positive. A low false positive value in the WMH will be recommended to the reference value of the block boundary or window hash. This method can be used as a solution for addressing the problems of storage media size required on the network forensic activity.

Corresponding Author:

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

1. INTRODUCTION

Services development offered by the internet service provider increase highly, because most people and organizations rely on their daily communication needs on internet services. As a consequence, security threats increase. Data from the Internet Crime Complaint Center (IC3), the number of complaints received in the year 2010 as many as 303.809 cases [1]. Research conducted by Chung et al [2] reported that in Taiwan 90 percent of Internet users conduct violations, and 71 percent of detected threats come from insiders. The need for internet crime reconstruction to find perpetrators of crimes is increase.

Traceback analysis is a common technique used in the reconstruction process of internet crime. Log data is the basis of the data in an investigation. The size of log files ranges from a few megabytes to terabytes on small networks of high traffic on the daily network. The higher the internet crime rate, the greater the media storage, and the greatest storage space for later analysis (post mortem analysis). The main data source to be used in reconstruction process is log data. Automatic data logging is the process of collecting and recording data from sensors for the purpose of archiving or analysis [3]. A log data must be organized and analyzed in order to reveal an information. Routine log correlation analysis is beneficial for identifying security incidents, policy, violations and problems. It is important to manage log data in order to monitoring network activity efficiently [3].

Header and Payload containing important information should be considered to create a profile on the network attacks [4]. Traditionally, there are two approaches used to identify the IP network traffic, namely Transmission Control Protocol (TCP) and / or User Datagram Protocol (UDP). The second approach includes a more sophisticated technique, based on deep packet inspection of TCP or UDP payload to search for a particular signature [5]. The weak point of capturing all the traffic and analyzing each type of crime is its use of large media storage. Traffic capture in 1 day in WAN scale with the number of hosts 300 requires as much a storage capacity as 1 TB [6]. Others have reported that the data at 1 Gbps link requires 10 TBytes daily to keep all traffic (storing full packet traces) [7]. Payload attribution is an important element in the network forensic analysis. Payload attribution provides packet transmission history and excerpt of possible packet payload [8]. The purpose of the payload attribution is to improve efficiency of traffic related to the efficiency of the media storage. To identify where an attack originated, how it propagated, and what computer(s) and person(s) are responsible, traceback and attribution are performed during or after a cyber violation [9].The aim of network forensic activity is to determine table routing from a victimized network or system to the point of attack origination or the person who is responsible. However, many common rule-based Intrusion Detection Systems (IDSs) use both packet header and payload data for event detection. This is also called Deep Packet Inspection (DPI) [10].

Attacks on computer networks can also be detected by payload or often called the payload detection rule. Content, nocase, rawbytes and offset option is a small part used to detect anomaly traffic with payload as the keyword. Winnowing Multi Hashing (WMH) is a method to improve the efficiency of data storage by reducing the payload size [8]. This study will provide on how the implementation of IDS with payload and implement rule-based WMH as traffic efficiency solutions. The remainder of this paper is organized as follow, in section 2, a brief overview of related work is presented. In section 3, we introduce the Payload Attribution System, Bloom Filter, Rabin fingerprinting, and Winnowing. The benefits of payload attribution can be used for Network Forensic using Payload Attribution. This section also discussed theories about fingerprinting and document fingerprinting using of *k*-grams. In section 4 we discuss the Winnowing Multi Hashing algorithm (WMH), which consists of Generating Fingerprinting and False Positive, section 5 displays the results of the experiment which include Network Topology, Calculation of WMH and Calculation of False Positive. Finally, in section 6 we will get a conclusion.

2. RELATED WORK

Network anomaly detection can be seen from its payload. As most of network anomaly detections based on packet headers, while payload is neglected [11]. Anomaly detection methods have been performing on each application level attacks, later the concept of the keywords in the payload associated with the detection of attacks is developed. Network traffic classification is to identify the protocols or types of protocols in the network traffic. In particular, the identification of network traffic with high resource consumption, such as peer-to-peer (P2P) traffic, represents a great concern for Internet Service Providers (ISP) and Network Managers[12]. Studies have been conducted to extract useful information in the payload to detect an attack. This method has a better detection mechanism . Some advantages of the payload-based IDS are [13]:

1. Generality of the system: the system is able to detect various types of applications and protocols.
2. Incremental profiling: system that accommodates changes in profile, which proactively detects changes in communicating patterns to the network activities which always changing.
3. Low false alarm rate: false alarm is very important in anomaly detection systems. Accuracy in detecting truths anomalous event is the main measure of reliability of an IDS.
4. Resistance to mimicry attack: mimicry attack is an attack by imitating the normal system, in order not to be detected by IDS. Payload-based IDS system is able to detect this type of attack.
5. Efficiency to-operate in the high bandwidth environment: high speed and high bandwidth networks are major requirements in the networks. Anomaly detection mechanism should be able to observe and analyze all traffic passing quickly and efficiently.
6. Unsupervised learning: IDS requires a training phase to update the rules or profile in mimicking an anomaly activity.

IP traceback can be used to find the origins and paths of the attacking traffic. There are two IP traceback techniques,namely Deterministic Packet Marking (DPM) and Probabilistic Packet Marking (PPM) [14]. The main problem with IP traceback used on the Internet is the difficulty of deployment. Incrementally deployable approach based on a sample flow for IP traceback technique ,known as Trace Sample, can be an option. IP traceback approach as part of network forensic activities can be divided into five groups: [15]

a. Link Testing
   Link testing focuses on the hop by hop IP traceback investigation to the router closest to the victim. The technique is repeated gradually to the establish router from one to another router to the source of attacked router [14].
b. The packets can be sent back to the source address or to the sender router. ICMP is the primary protocol that acts to mediate between the router with the host destination. Using the ICMP protocol is possible to trace the source of the attacker.
c. Logging

Logging traceback method is very efficient and effective in the reconstruction process to find out the source of the attacker. SPIE (Source Path Isolation Engine) developed related to the efficiency of the storage media, for many packages must be stored and analyzed. This package only took hashing and digest package [16].

d. Packet marking
To implement the IP Traceback services, it is important to allocate enough space in the IP packet header, so that the space can be used to record the transversed path of a package. For example, each router, in addition to packet forwarding and routing functions, the router also adds its ID in the packet header.

e. Hop Count Distance
Hop count distance is calculated based on packet signatures and determined hop count value, to predict the source of the attacker tough the IP source had been spoofing. This method uses the value of the TTL (time to live) hop count to trace the source of attacker. Traceback process through a router rated decremental -1 to the router closest to the victim.

Alternatively, Ponec et al. [8] proposed a number of attribution techniques, namely, the VBS, WBS, and WMH. These methods compare data reduction ratio and accuracy rates. This technique was developed to overcome the problem of offset collision in Bloom Filter on previous attribution technique. Rabin fingerprint technique is used to determine the block boundaries in a VBS method on its own window. The block boundary is set when the fingerprint mod m is equal to zero, where m can be any arbitrary value. The Winnowing technique is used in WBS, to determine the boundaries [5]. Recent developments on WMH incorporate the concept of multiple instances of WBS with difference parameters and reduction of false positives. Hashem et al. developed a payload attribution techniques by modifying to the Bloom Filter with Character Multi Dependent Bloom Filter (CMBF) [7]. The works of developed CMBF are CMBF employs 256 distinct bloom filters in order to create a one-to-one mapping for each possible way of bytes to a unique bloom filter. CMBF uses fingerprint modulo (q) in the fingerprint calculation, which maps each string to a byte value (class) between 0 and q-1.

3. PAYLOAD ATTRIBUTION SYSTEM

Payload attribution system (PAS) is a partial quoting payload (excerpt) technique to drive traffic to efficient media storage. Payload and query processing are two main tasks in the PAS. In payload processing, all traffic that passes on the network is tested according to the existing rules, as well as some of the data stored in permanent storage. Filtering technique is often used in the process of capturing in example on HTTP traffic only. This data is stored later in the archive unit, which has two timestamps, the start and end time interval collecting data. FlowIDs information includes source and destination IP addresses are also required for the identification of data packets. The second task of PAS, querying processing, is an important part to retrieve all to the storage unit, then performed an excerpt for each flowID and report its activity. Several techniques have been developed such as the Hierarchical Bloom Filter (HBF), Variable Block Shigling (VBS), Winnowing Block Shingling (WBS) and Winnowing Multi Hashing (WMH) [9]. Three basic concepts in payload attribution are:

3.1 Bloom filter

Bloom filter, by Broder and Mitzenmatcher [17], are developed for the purpose of the network and other applications, and formulated as equation (1).

$$\propto = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^{k} \approx \left(1 - e^{-kn/m}\right)^{k} \qquad \text{.........................(1)}$$

Bloom filter which is implemented on the network, constitute the division of payload on blocks according to hierarchy or existed level. The advantage of this technique is easily implemented by simply determining byte block size = s, then each payload is divided into blocks, then at the next level coupled between one block to another, which is followed by the Offset value as identifier of a content payload. The drawback of this technique is not capable to deal with double block offset or often referred as offset collision. Additionally, high false positives value is also a factor that should be further investigated [8]. Hashing , a fundamental tool in digital forensic analysis, is used both to ensure the data integrity and to identify data objects. Md5 bloom revised enables network forensic activities become more efficient both in terms of scalability and also originality of evidence [18].

3.2 Rabin Fingerprinting.

This method gives a probability value of checksums as markers on two different objects which has same fingerprinting values.

$$s(x) = s_1 x^{N-1} + s_1 x^{N-2} + \cdots + sN \quad \text{................(2)}$$

Equation (2) was developed for http traffic or web caching on a computer network [19]. Basically this method is quite well to set a fixed block size with a variable, so that the payload will follow the capacity of the appointed block. The susceptibility of this method is that there were possible to appear many blocks but in a small size or a block in a very large size.

3.3 Winnowing

Winnowing [20] is one method to modify the Rabin fingerprinting technique to detect all or part of the necessary documents. Each sequence of characters will be stored in the array storage. This method assumed to be most appropriate to use to block boundaries in the payload attribution. The first item will be stored in a hash c1, c2, ... cv, and the second item will be stored in a hash c2, c3 ... cv +1 and so on where ci is the value of the characters in the document with Ω bytes, for i = 1. Ω will make on a sliding windows with *w* size to the array and choose the hash for each windows. Markers are used to determine hashing as in Rabin's fingerprinting. The weakness of this method is the lack of a good control of the block and the maximum size of the payload. The advantages of this method are the low false positive value, and considered most suitable for payload attribution in network forensics.

3.4 Network Forensic Using Payload Attribution

Network forensics is the process of capturing, recording, and analysis of network data traffic which lead to a security threats in computer networks [21]. Meanwhile, according to Almulhem and Issa [22], network forensics is developing a network security model that focuses on the capture, recording, and analysis of network traffic for the purposes of investigation. The concept of network forensics deals with the data found in computer networks and Analysis of network traffic such as data logging via a firewall or intrusion detection systems on devices like routers. Network Forensic Analysis Tool is a tool used in this activity. The importance of this work is to present an overview of network forensics, covering tools, models and frameworks implementation process, which will be very useful for practitioners of network computer security [23]Traceback process was conducted to determine the identity of the attacker in order to be used as a source of guidance on the investigation process. The increasing number of Internet users and the higher number of media crimes, the extraction traffic or efficiency in storage media is a must. Data were obtained from the data log as the activity history for each user then compressed using the WMH. After this compression process is completed, traceback analysis is then performed to obtain attacker profile, which is synchronized with the header as a provider of information a data packet.

3.5 Fingerprinting

Digital watermarking offers additional form of protection to a broader scope. Watermarking will include some information in the document, such as audio tracks, still images, video streaming or text. Fingerprinting is one of the digital watermarking techniques as a tool for copyright protection of the document. Unique mark added to each copy of the document to then be used as a tracking device by the owner or distributor of pertinent documents Two fingerprinting techniques used today are symmetric and Asymmetric techniques. Most users use symmetric techniques. On symmetric techniques, distributor and fingerprinting holders have equal access to the document fingerprint. It is difficult to determine who is responsible if there is case of fraud. Asymmetric fingerprinting techniques make different fingerprint, allow the users who only know about their fingerprint data. Distributor for the first time will do public key encryption. Then the user will do the description with the secret key.After the deal is done, users have a uniquely, and tied to him, fingerprinted copy of the data. The distributor does not get this copy. From the public key gained in the key exchange, the distributor cannot create a copy identical to the one that the user has.

3.6 Document Fingerprinting

Document fingerprinting is a technique to detect full and partial copies between documents. It works by storing a small scale (small sketch) that is representative to a set of numbers. Benchmarking sketch between two documents will provide information document substantially overlap or not. A digital document fingerprinting scheme is a particular position marked on a document. Fingerprinting algorithm will match the same number of documents that are in the specified position.

3.7 Document Fingerprinting using hashes of k-grams

One technique to document hash (checksum) in the detection of duplicates in document is by comparing two similar documents. Checksum value on a document will represent all of your content, so by comparing the value will provide plagiarism information or not. This technique is then improved to be more perfect with the use of *k*-grams of hash functions. This technique uses a specific part of a document as fingerprinting used to create hashing. Particular part is a *k*-grams.

4. RESEARCH METHOD

Winnowing Multi Hashing method is one of the many methods that can be applied to streamline payload [8]. One advantage of this method is to use an efficient storage medium than the previous method. In particular, this method shows the best technique for the selection ofthe Payload block boundary. Winnowing algorithm is derived from document fingerprinting technique, by eliminating documents overlapping n-grams (sequences of N words) and then performs conversion of every n-gram hash value. The algorithm selects the smallest hash value from each overlapping window of X sequential hashes as document fingerprints [24]. Document comparison is then reduced to find exact matches in the sets of fingerprints to determine the fingerprints not two things first, overlapping documents will be guaranteed to have intersecting sets of second fingerprints, the sets of fingerprints should be small enough to permit scaling the task to large number of documents. The following session will explain how the work of Winnowing algorithm with a case study for selecting fingerprints from hashes of k-grams [25]. We give an upper bound on the performance of Winnowing, expressed as a trade-off between the number of fingerprints that must be selected and the shortest match that we are guaranteed to detect.

To find substring matches in a document there are two things to note:

1. *If there is a substring match at least as long as the guarantee threshold, t, then this match is detected, and*
2. *We do not detect any matches shorter than the noise threshold, k.*

The constants t and $k \le t$ are chosen by the user. We avoid matching strings below the noise threshold by considering only hashes of k-grams. The larger k is, the more confident we can be that matches between documents are not coincidental. On the other hand, larger values of k also limit the sensitivity to reordering of document contents, as we cannot detect the relocation of any substring of length less than *k*.

The following example will explain the Winnowing algorithm with parameter values specified by the user k is worth 5.

A do learn learn learn, a do learn learn

(a) Some text.
adolearnlearnlearnadolearnlearn

(b) The text with irrelevant features removed.
adole dolea olear learn earnl arnle rnlea nlear learn earnl arnle rnlea nlear learn earna arnad rnado nadol adole dolea olear learn earnl arnle rnlea nlear learn

(c) The sequence of 5-grams derived from the text.
98 70 42 17 88 50 12 78 17 88 50 12 78 17 23 7 13 20 98 70 42 17 88 50 12 78 17

(d) A hypothetical sequence of hashes of the 5-grams. Define a window of size w to be w consecutive hashes of k-grams in a document (w is a parameter set by the user). By selecting at least one fingerprint from every window the algorithm limits the maximum gap between fingerprints. In fact, the algorithm is guaranteed to detect at least one k-gram in any shared substring of length at least $w + k - 1$. Given a sequence of hashes h1…….hn, if $n > t - k$, then at least one of the hi must be chosen to guarantee detection of all matches of length at least t. This suggests the following simple approach.

4.1 Generating Fingerprints

(a) Minimun hash values are selected from each windows hashes of length 5
(98 70 42 **17** 88)(70 42 17 88 50)(42 17 88 50 12)(17 88 50 **12** 78)(88 50 12 78 17)(50 12 78 1788)
(12 78 **17** 88 50)(78 17 88 50 12)(17 88 50 **12** 78)(88 50 12 78 17)(50 12 78 17 23)(12 78 17 23 **7**)
(78 17 23 7 13)(17 23 7 13 20)(23 7 13 20 98)(7 **13** 20 98 70)(13 **20** 98 70 42)(20 98 70 42 **17**)
(98 70 42 17 88)(70 42 17 88 50)(42 17 88 50 12)(17 88 50 **12** 78)(88 50 12 78 17)

Selected fingerprint is

17 2 17 12 7 13 20 17 12

(b) Fingerprints paired with 0-base positional information.

[17,4] [12,7] [17,9] [12,12] [13,17] [20,18] [17,22] [12,25]

Choose the minimum hash value for each of the windows as the following equation :

$$h = (h_1, h_2, \ldots\ldots h_x)(h_2, h_3, \ldots. h_{x+1}), (h_3, h_{4,\ldots\ldots} h_{x+2}) \qquad \ldots\ldots\ldots\ldots\ldots(3)$$

The forming of fingerprinting is derived from a window of X size to be X consecutive hashes (h) (X is the parameters set by the publisher). From instructing window, choose the value of h min with the smallest value. If there were two values has the same grade, choose the right most position. As stated before, choosing the h min, the h min in one window is very likely to stay the minimum hash in adjacent windows. Thus many overlapping windows select the same hash, and the number of fingerprints selected is much smaller.

4.2 False Positive

False positives reported the abuse while none abuse had happened . In the context of signature-based detection, false positive occurs when the signature has compatibility with normal activity. This could be the result of two scenarios. The first signature is not specific on the search of an attack signature or the second condition is too specific but normal traffic and malicious match with the signature. False Positive (FP) can be calculated by the relation of two equations

$$\left(1-\left[1-\frac{1}{m}\right]^{kn}\right)^{k} \approx \left(1-e^{-kn/m}\right)^{k} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Where m = payload length, N is number of inserted items, and k is the number hash fuction, k = ln2. m / n, and e = Limitation While the value of e is:

$$\lim_{x\to\infty}\left(1-\frac{1}{x}\right)^{-x} = e \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

4.3 Research Design

This study will try Winnowing algorithm implementation on network forensic activity by capturing payload as a data input. Research design can be seen in Figure 1.

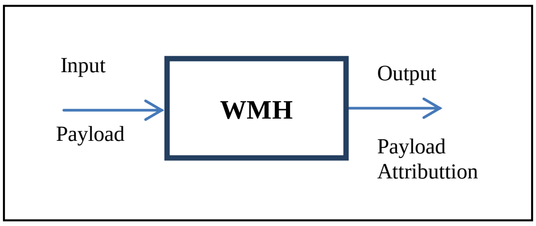


Figure 1.Research Design

In Figure 1 the input received is a payload captured by a content rules-based IDS device. From input received, extracted with WMH in order to get an efficient payload, named the payload attribution.

4.4 Network Topology

Figure 2, the architecture in the forensic process that consist of four routers. Four routers represent router network id. The function of a router is as a liaison between two or more networks to carry data from one network to another [8].

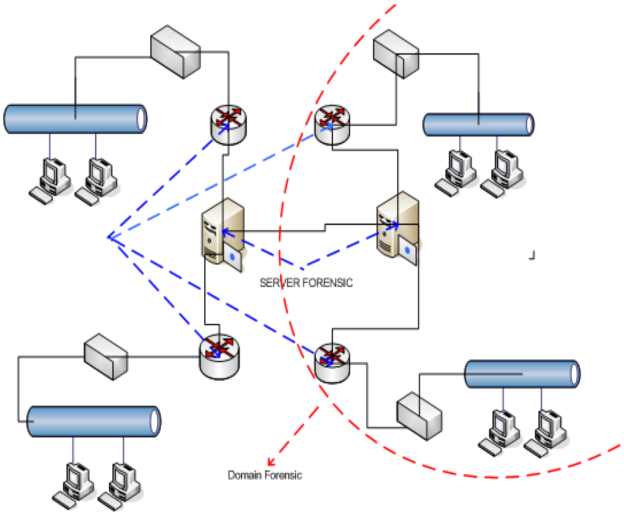


Figure 2 Network Topology

The routers default will run the routing protocol and the forwarding of data packets, so that packets arrive at the destination. In a process of network forensics, forensics server function is to capture traffic by using of intrusion detection rules based on payload system.

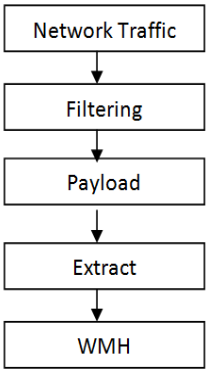
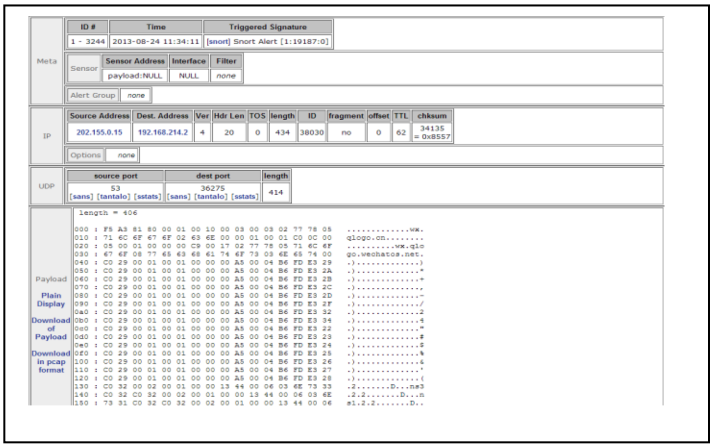


Figure 3 Flow Capturing Process

Figure 3 displays data capturing process. It will be explained as follows:

- Network Traffic
  The first step is to capture existing network traffic. Capturing the network traffic can be done by using the IDS Snort or wireshark. This phase will capture traffic anomalies, according to the rules on payload option. To classify the network traffic, regular expression signatures applied. There are 4 kinds of common approaches, include, pre-processing to extract the payload application session, tokenization to find common substrings and incorporate position constraints, multiple sequence alignment to find common subsequences, and signature construction to transform the results into regular expressions [26].
- Filtering
  Because so many protocols become targets of attack, it is necessary to filtering. This research will focus on attacks that use TCP transport. On developing DPI methods, Bloom Filters techniques used to match the string to avoid a bottleneck [27]. Pattern storage efficiency according to the specified rule is also an important concern in building an NIDS. Database compression pattern matching techniques to snort into the bit vector will further reduce the complexity of the hardware specifications [28]. Protocols such as HTTP, TCP, UDP, and so on are results captured on a Snort . Filtering process is a process to filter HTTP-based protocol.

- Payload
  The payload is the actual data received by the destination. While the Header, as the profile packet, is discarded when it reaches the destination. In this experiment, the payload applied is the payload anomaly, detected by the Snort payload detection rule. Some rules frequently used are content, rawbyte, depth, offset and others. Sample results capturing payload based on snort IDS shown in Figure 4. It is appear on figure 4 that the payload length of actual data recording is 406 byte.



ID #	Time	Triggered Signature
1 - 3244	2013-08-24 11:34:11	[snort] Snort Alert [1:19187:0]

Sensor	Sensor Address	Interface	Filter
	payload:NULL	NULL	none

Alert Group	none

Source Address	Dest. Address	Ver	Hdr Len	TOS	length	ID	fragment	offset	TTL	chksum
202.155.0.15	192.168.214.2	4	20	0	434	38030	no	0	62	34135 = 0x8557

Options	none

source port	dest port	length
53 [sans] [tantalo] [sstats]	36275 [sans] [tantalo] [sstats]	414

```
 length = 406

000 : F5 A3 81 80 00 01 00 10 00 03 00 03 02 77 78 05   .............wx.
010 : 71 6C 6F 67 6F 02 63 6E 00 00 01 00 01 C0 0C 00   qlogo.cn........
020 : 05 00 01 00 00 00 C9 00 17 02 77 78 05 71 6C 6F   ..........wx.qlo
030 : 67 6F 08 77 65 63 68 61 74 6F 73 03 6E 65 74 00   go.wechatos.net.
040 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 29   .).............)
050 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 2A   .).............*
060 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 2B   .).............+
070 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 2C   .).............,
080 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 2D   .).............-
090 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 2F   .)............./
0a0 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 32   .).............2
0b0 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 34   .).............4
0c0 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 22   .)............."
0d0 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 23   .).............#
0e0 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 24   .).............$
0f0 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 25   .).............%
100 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 26   .).............&
110 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 27   .).............'
120 : C0 29 00 01 00 01 00 00 00 A5 00 04 B6 FD E3 28   .).............(
130 : C0 32 00 02 00 01 00 00 13 44 00 06 03 6E 73 33   .2.......D...ns3
140 : C0 32 C0 32 00 02 00 01 00 00 13 44 00 06 03 6E   .2.2.......D...n
150 : 73 31 C0 32 C0 32 00 02 00 01 00 00 13 44 00 06   s1.2.2.......D..
```

Figure 4 Snapshot of a Payload

- Extract
  Extracting is a process to take bits of payload. Payload bits can be either binary or hex. The hex payload bits is used in this study. Extraction process detached from the process of catching traffic. This process can be done offline, but in the future the system will be assisted with the extraction tools so that the process can be done faster.
- Winnowing Multi Hashing (WMH)
  Winnowing Multi Hashing is a method used for payload attribution.
  The winnowing algorithm is an instance of document fingerprinting, which cuts the document into all overlapping n-grams (sequence of N words) and then translates each n-gram into a hash value.

5. RESULTS AND ANALYSIS

This section will discuss the results of the research. There are two important things to be done, the first one is the application of WMH in the payload that has been captured by the IDS, then the second one is to calculate the value of false positives will be used as a reference in determining the window size used**.**

5.1 Calculation of WMH

Results trial of WMH method in this study as follows.

(a) Byte payload obtained from capturing process using original Snort IDS in the form of hexa-decimal.
66:39:61:0d:0a:ec:7d:09:77:db:c6:d2:e5:5f:c1:d1:3b:e3:93:cc:a7:85:00:08:80:74:c6:ce:d1:6a:ed:92:45:2d:b6:a2:0c:0f:48:82

(b) The next step is to define Window hash value. This study will try window hash=5 which is set by the user. The payload is divided into blocks with a block size is 5.

(66 39 61 0d **0a**) (39 61 0d 0a ec) (61 0d 0a ec 7d) (0d 0a ec 7d 09) (0a ec 7d **09** 77) (ec 7d 09 77 db)
(7d 09 77 db c6) (09 **77** db c6 d2) (77 db **c6** d2 e5) (db c6 d2 e5 5f) (c6 d2 e5 **5f**c1) (d2 e5 5f c1 d1)
(e5 5f c1 d1 3b) (5f c1 d1 **3b** e3) (c1 d1 3b e3 93) (d1 3b e3 93 cc)(3b e3 **93** cc a7) (e3 93 cc a7 85)
(93 cc a7 85 **00**) (cc a7 85 00 08) (a7 85 00 08 80) (85 00 08 80 74) (00 **08** 80 74 c6) (08 80 **74** c6 ce)
(80 74 c6 ce d1) (74 c6 ce d1 **6a**) (c6 ce d1 6a ed) (ce d1 6a ed 92) (d1 6a ed 92 45) (6a ed 92 **45** 2d)
(ed 92 45 2d b6) (92 45 2d b6 a2) (45 2d b6 a2 **0c**) (2d b6 a2 0c 0f) (b6 a2 0c 0f 48) (a2 0c 0f 48 82)

(c) The mapping process window assuming block hash=5, the value of the extracted payload fingerprinting can be seen as
0a 09 77 c6 5f 3b 93 00 08 74 6a 45 0c

(d) The determination of the position or address of fingerprinting extraction results are follows

[0a, 4] [09, 7] [77.8] [c6, 10] [5f, 13] [3b, 16] [93, 18] [00, 22] [08.23] [74.25] [6a, 29] [45.32]
[0c, 26]

Explanation of the data above, i.e. [04, 4] : 04 is the payload fingerprinting, while 4 is the serial number or address fingerprinting.

5.2 Calculation of False Positive

Calculation of the value of the false positive tolerance e according to equation 3 above assisted with limit values that are often used as Table 1.

Table 1. Value limitation

X	$\left(1-\frac{1}{x}\right)^{-x}$
4	3.160494
16	2.808404
64	2.739827
256	2.739827
1024	2.719610
4096	2.718614
16384	2.718365
65536	2.718303
262144	2.718287
1048576	2.718283
4194304	2.718282

The processes of calculating the value of false positive in equation 2 are follows. On Fingerprint [04.4], the value of M=04 is payload) and N=4 is the sequence number. K value is multiplied LN2 M divided by n then the calculation results obtained 0.693. Having obtained the value of K, then in calculating the false positive values such as 3, the value of false positive (FP) =0.0095 using constant value e=2,808 according to Table 1.

[04, 0] is the first fingerprint obtained accompanied by the fingerprint position in the overall hash value obtained. e is the letter to state limits. Limit used freely determined by the user. In this example e approximation used is 2.808404 obtained from the limit equation with x = 16.

Table 2 Value of False Positive

	Value of false positive			
Finger Print	Window Hash =3	Window hash =5	Window Hash = 40	Window Hash =1
[0a, 4]		0,0095		
[39,2]	0,0001			
[0a, 4]			0,6	
[66,1]				0.0000

Table 2 shows the calculation of false positive results on windows hashing = 5 with fingerprinting position [04.4] is 0.0095. Experiments on windows hashing size = 3 with the fingerprinting [39.3] false positive results obtained = 0.0001. Experiments on the window hash=40 with fingerprinting [04.4] is 0.6. Recent experiments with fingerprinting [66.1] with window hashing=1, obtained the lowest false positive value, 0.0000.

CONCLUSION

We get four conclusions in this paper, related to the implementation of Winnowing Multi Hashing on payload attribution activity.

1. WMH method can be used to carry the payload attribution. WMH Method can be used as a solution for addressing the problems of storage media size required on the network forensic activity.
2. Calculation of false positives can be used as a determinant of the amount of the value of most appropriate *k* gram. False positive values using windows hashing = 1 is zero. It means to no payload attribution.

3. The smaller the windows hash value, the smaller false positive value obtained. The experiment shows that window hashing=3 is a top priority in conducting payload attribution.
4. This method can be used for the efficiency of traffic on internet crime reconstruction. But in subsequent studies, the evidences originality should be tested.

AUTHOR CONTRIBUTIONS

On this study, the contribution of the authors is to implement algorithms Winnowing Multi Hashing on the actual data packet (payload) in computer networks. Inefficiency problems related to storage traffic in Internet crime investigation activities (network forensic) can be solved by this method. So far, this method is used to check the duplicate content of the text.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their constructive comments which has led to an improved version of the manuscript**.**

REFERENCES

[I] Internet Crime Complaint Center (IC3)., *2010 https://www.ic3.gov/complaint/default.aspx*.

[2] W.Chung, et al., "Fighting cybercrime: a review and the Taiwan experience," Elsevier B.V. All rights reserved, 2004

[3] A. Madani, *et al*., "Log Management comprehensive architecture in Security Operation Center (SOC)," *International Conference on Computational Aspects of Social Networks (CASoN)*, 2011

[4] F.M. Cheema, *et al.,"*Comparative Evaluation of Header vs. Payload based Network Anomaly Detectors," *Proceedings of the World Congress on Engineering*, Vol I- 3, 2009, London, U.K

[5] R. Ashamarri and A.N.Z Heywood," Can encrypted traffic be identified without port numbers, IP addresses and payload inspection?," *Elsevier B.V,* All rights reserved, 2010.

[6] P.Giura and N.Memon, "An efficient storage infrastructure for network forensics and monitoring," Proceeding *RAID'10 Proceedings of the 13th international conference on Recent advances in intrusion detection*, Pages 277-296 Springer-Verlag Berlin, Heidelberg, 2010.

[7] M. H. Haghighat, *et al*., "Payload Attribution via Character Dependent Multi-Bloom Filters," *IEEE Transactions on Information Forensic and Security*, Vol. 8, no. 5, May 2013.

[8] 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.

[9] A. Lazzez, "A Survey about Network Forensics Tools," *International Journal of Computer and Information Technology,* (ISSN: 2279 – 0764)Volume 2– Issue 1, January 2013.

[10] T. Limmer and F. Dressler, "Improving the Performance of Intrusion Detection using Dialog-based Payload Aggregation *",IEEE Global Internet Symposium (GI) 2011,* at IEEE INFOCOM 2011.

[11] H.Tian, "An Incrementally Deployable Flow-Based Scheme for IP Traceback," *IEEE Communication Society,* ISSN 1089-7798, 2012.

[12] J.Camacho,et al., "A generalizable dynamic flow pairing method for traffic Classification," *Elsevier B.V.* All rights reserved, 2013.

[13] S.A. Thorat, *et al*., "Payload Content based Network Anomaly Detection,", ©2008 IEEE

[14] C.Lokesh, *et al*.," ETM: a Novel Efficient Traceback Method for DDoS Attacks," *International Journal of Computer Science and Management Research,*Vol 1 Issue 3, October 2012.

[15] S. Savage, *et al*., "Practical network support for IP traceback," *IEEE/ACM Transaction on Networking*, " Vol. 9, No. 3, June 2001.

[16] A. C. Snoeren, *et al*., "Single-Packet IP traceback," *ACM,* copyright 2001.

[17] A. Broder and M. Mitzenmatcher," Network Applications of Bloom filters: A survey, " *In Proceedings of the Annual Allerton Conference on Communication Control and Computing*, SIAM, Philadelphia, 2002.

[18] V. Roussev, *et al.,"* md5bloom: Forensic filesystem hashing revisited," *DFRWS,* Published by Elsevier Ltd, 2006

[19] S. Rhea, *et al*., "Value-based Web caching*," In Proceedings of the 12thInternational World Wide Web Conference. ACM*, New York, 2003.

[20] S. Schleimer, 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.

[21] M. Ranun and J,Arcus," *Network Forensic and Traffic Monitoring Computer Security*," *Computer Security Journal*, Volume XII, November 1997.

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

[23] E. S. Pilli, *et al*., "Network forensic frameworks: Survey and research challenges," *Elsevier Ltd* , All rights reserved, 2010

[24] S.M. Darwish, "New system to fingerprint extensible markup language documents using winnowing theory,*" IET Signal Process,* Vol. 6, Iss. 4, pp. 348–357 & The Institution of Engineering and Technology, 2012

[25] N. Elbegbayan, 2005, "Winnowing, a Document Fingerprinting Algorithm," *Department of Computer Science Linkoping University*, 2005

[26] Y.Wang, *et al*.," Generating regular expression signatures for network traffic classification in trusted network management," *Elsevier Ltd,* All rights reserved, 2011.

[27] K.Huang, et al., “Accelerating the bit-split string matching algorithm using Bloom filters,” *Elsevier Ltd*, 2010.
[28] N. B. Guinde and S. G. Ziavras,” Efficient hardware support for pattern matching in network intrusion detection,” *Elsevier Ltd*, All rights reserved, 2010.

BIOGRAPHIES 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 . His research 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. He received his S1 degree in Statistics from Gadjah Mada 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