A New Logging-based IP Traceback Approach using Data Mining Techniques

Ho-Seok Kang and Sung-Ryul Kim*
Internet & Multimedia Engineering, Konkuk University, Seoul, Republic of Korea
hsriverv@gmail.com, kimsr@konuk.ac.kr

Abstract

IP Traceback is a way to search for sources of damage to the network or host computer. IP Traceback method consists of reactive and proactive methods, and the proactive method induces a serious storage overhead. However, a system capable of solving these problems through cluster-based mass storage, digestible packets and hierarchical collections was designed. It not only performs traceback but also communicates with analysis data of other security systems by using the logging methods. It is capable of performing an effective traceback operation by using data mining in order to perform vast amount of traceback operation with the use of massive data. In addition, the results can be used as basic data to generate new rules for intrusion detection systems.

Keywords: IP Traceback, logging-based approach, data mining

1 Introduction

With the development of computer and internet, various methods to interfere with computer work have been developed and advanced. Various forms of attacks such as computer virus, computer hacking, computer worm, and DoS (Denial of Service) attack have emerged. Recently, these attack methods have been further developed and has become more complex. Moreover, security managers are not aware of the attack during these complex attack methods are carried out.

In general, methods to protect against these network attacks undergo a two-stage process. The first is to detect an attack and report it to the manager while it is being made. The second is a traceback stage to trace the source of the attack. In the first stage, real-time intrusion and DDoS (Distributed Denial of Service) attacks are blocked. In the second stage, the traceback detects the intrusion path by searching for the host or network that launched an attack in order to prepare for future attacks.

Among them, traceback method consists of reactive and proactive methods [13, 10]. The reactive method performs the traceback during an attack. The proactive method performs the traceback by analysis after the attacks. In the proactive method, the most generally used methods are to mark packets and store all the packets. We call these methods marking-based and logging based method [12]. Although the marking-based method can trace back only the DDoS attacks, it can save storage compared to the method to store all packets. On the other hand, logging-based method has the advantage that not only DDoS attack but also a single packet can be traced back by storing only important information of the packet. In addition, there is a hybrid method made by mixing these two methods. However, hybrid method has the advantages and disadvantages of both methods.

The marking-based method can only traceback the DDoS attacks with large-scale traffic. Therefore, the logging-based method can be better when sufficient storage space is provided. Furthermore, it is possible to detect a new malware or create a new policy by using the stored data. However, because the

Journal of Internet Services and Information Security (JISIS), volume: 3, number: 3/4, pp. 72-80

*Corresponding author: 120 Neungdong-ro, Gwanjin-gu, Konkuk University, Seoul 143-701, Republic of Korea, Tel: +82-(0)24504134, Web: http://aislab.konkuk.ac.kr

method to store the information requires to store vast amounts of traffic data, methods such as packet summary information or alternative storage are frequently used. In this paper, we propose a security network framework for enterprise security management [2]. In addition, the traffic data is analyzed and stored by data mining process. Thus, the data mining information is used to find various traceback paths easily.

Although this method uses the logging-based traceback method, it summarizes and stores only main information from large-scale data by using data mining. This reduces an enormous amount of calculation used for traceback and at the same time helps reduce the storage overhead by securing storage space of the cluster DB and distributed position. Moreover, this method is capable not only of DDoS traceback but also of single packet traceback.

The rest of this paper is organized as follow: Chapter 2 describes related research, Chapter 3 describes security management system framework proposed by us. Chapter 4 describes a simple scenario for traceback, and Chapter 5 draws a conclusion.

2 Related Work

The traceback method generally consists of reactive and proactive methods. When an attack starts, the reactive method traces the attacker from the victim by repeatedly querying to an upstream router until it finds out the attacker. However, this method cannot trace when the attack ceases. In the proactive method, a router stores information capable of tracing an attacker. Subsequently, it detects the attacker by analysis of the stored data. The proactive method can be roughly categorized into logging-based approaches, marking-based approaches, and hybrid approaches. The logging-based approaches can traceback both single packet and DDoS attack. However, its storage overhead is very high. Although the marking-based approaches use only small storage space, they can traceback only DDoS attacks. In order to compensate for the disadvantages of these two methods, the hybrid approaches have appeared.

Snoeren et al. [11] studied methods for IP traceback of hash-based single packets. This method called SPIE is the first logging-based approach that saved the storage space by using the method to store only packet summary. A bloom filter, a hash-based space-saving data structure, was used for the packet summary. However, SPIE rapidly runs out of space in high-speed network. T.Lee et al. [8] proposed to digest packet aggregation units instead of individual packets so as to reduce the digest table storage. However, that increases the false positives in constructing the attack graph. Chao Gong et al. [5] proposed PPM. PPM is a marking-based approach that traces only DDoS attack. B. Al-Duwairi et al. [1] proposed a hybrid approach called DLLT. The main idea of DLLT is to keep track of a subset of the routers that are involved in forwarding a certain packet by establishing a temporary link between them using distributed link list. However, DLLT method cannot trace single packets due to its probabilistic nature. Chao Gong et al. [6] proposed new hybrid approach called HIT. HIT can trace single packet. It records every packet's path fragments. Each router carries out both marking and logging. However, this method brings storage overhead by high-speed links of routers. Recently new hybrid approaches [12, 9] were proposed. However, these approaches have problems of storage overhead with high-speed links of routers.

Among the various methods, we designed a security system framework capable of tracing back by compensating the disadvantages of the logging-based approaches without correcting packets and then designed a traceback system based on this framework.

3 Framework of Hierarchical Security Management System

In order to realize a new logging-based method, we designed a hierarchical framework of security management system. The design principles consist of distributed and hierarchical storage, mining analysis and application of the analyzed content. Mass storage space is required for the framework because it depends on storage. However, it is difficult to handle the storage overhead by simply increasing storage space because network speed and traffic have increased dramatically in recent years. For this purpose, we need create cluster-based storage space capable of easy expansion and a function to extract only important information from traffic data for storage by using data mining techniques. Moreover, our framework is able to respond to lack of storage by distributing storage positions and hierarchically classifying them.

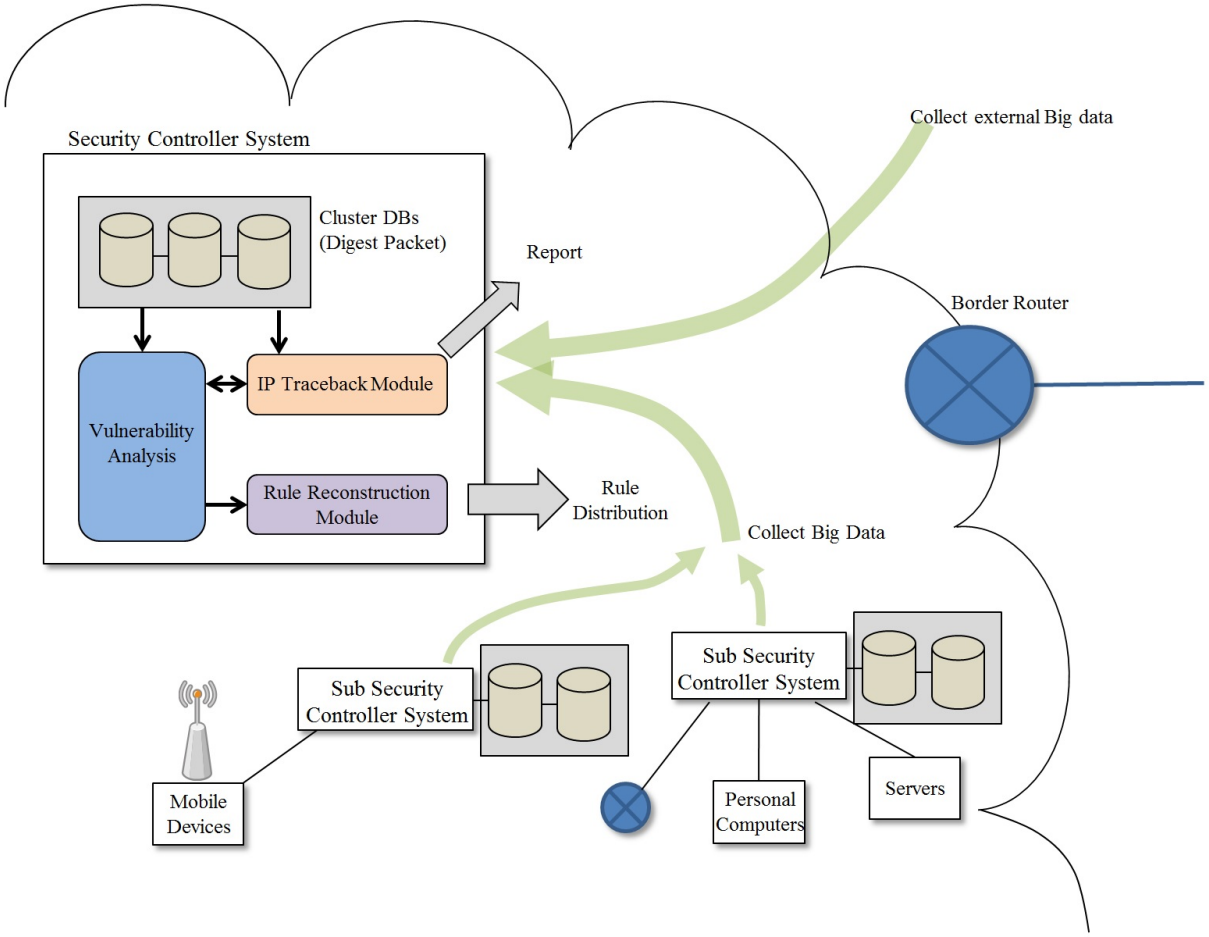

Figure 1: System diagram using logging-based approaches

Figure 1 is a diagram of integrated security system using these logging-based approaches. Based on this diagram, one can see that the system is divided into SCS (Security Controller System) and SSCS (Sub Security Controller System). These are the higher system SCS and lower system SSCS for hierarchical security management and collection. SCS periodically or selectively bring the traffic information stored in SSCS to store it. In addition, it brings the traffic information from the external SSCS or external routers to store it by classifying it into the desired form. Thus, the stored information is used for traceback using data mining or made into rules appropriate for IDS, firewall or lower SSCS for distribution.

In order to reduce packet information oriented to the cluster DB the central SCS, this framework used

the hierarchical collection system using SSCS. SSCS classified based on role or regional characteristics that are in the lower portion of SCS, and this SSCS collects the packets of the region. The classification broadly combines external networks, border routers, mobile routers, routers, PCs or servers of each region to manage them. For effective management, the collection can be divided by additionally placing SSCS in the lower portion of SSCS. Figure 2 shows a diagram of a system consisting of such a hierarchical collection.

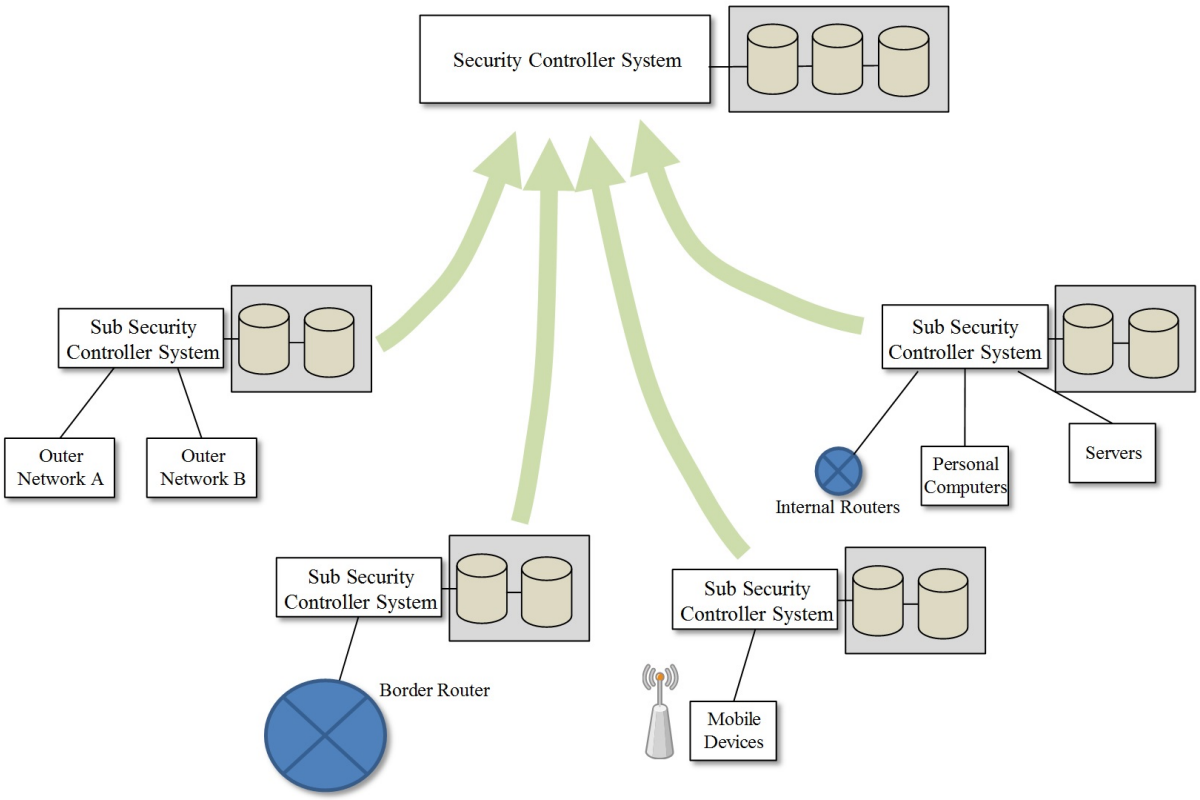

Figure 2: Diagram of hierarchical collection system

Subsequently, the traffic storage method based on each hierarchy level may be differentiated for more detailed analysis or management. When storing all traffic packets and transferring data to higher SSCS or SCS in lower SSCS, it is possible to reduce the number of packets by sampling or the storage volume in lower SSCS and SCS by transferring only summary information of each packet.

4 Mining-based IP Traceback

We describe the process to make system behavior and traceback path for a practical traceback on the basis of the integrated framework of security management system by using data mining. This is a proactive method that is performed after an attack and a method using mass storage data. Therefore, it takes a long time and many operating processes to analyze the stored data. In order to reduce these operations, data mining techniques are used.

4.1 Data Collection for Traceback

Only necessary information is extracted from the collected information by data mining and used for various analyses. First, in order to perform classification tasks by attack type for traceback using mining. The classification tasks by attack type apply packet pattern matching, a variety of classification algorithms and others [3, 4]. The classified data uses method to improve search speed by using the improved tree structure and save storage space [7]. Figure 3 shows the methods to perform and store classification tasks by attack type.

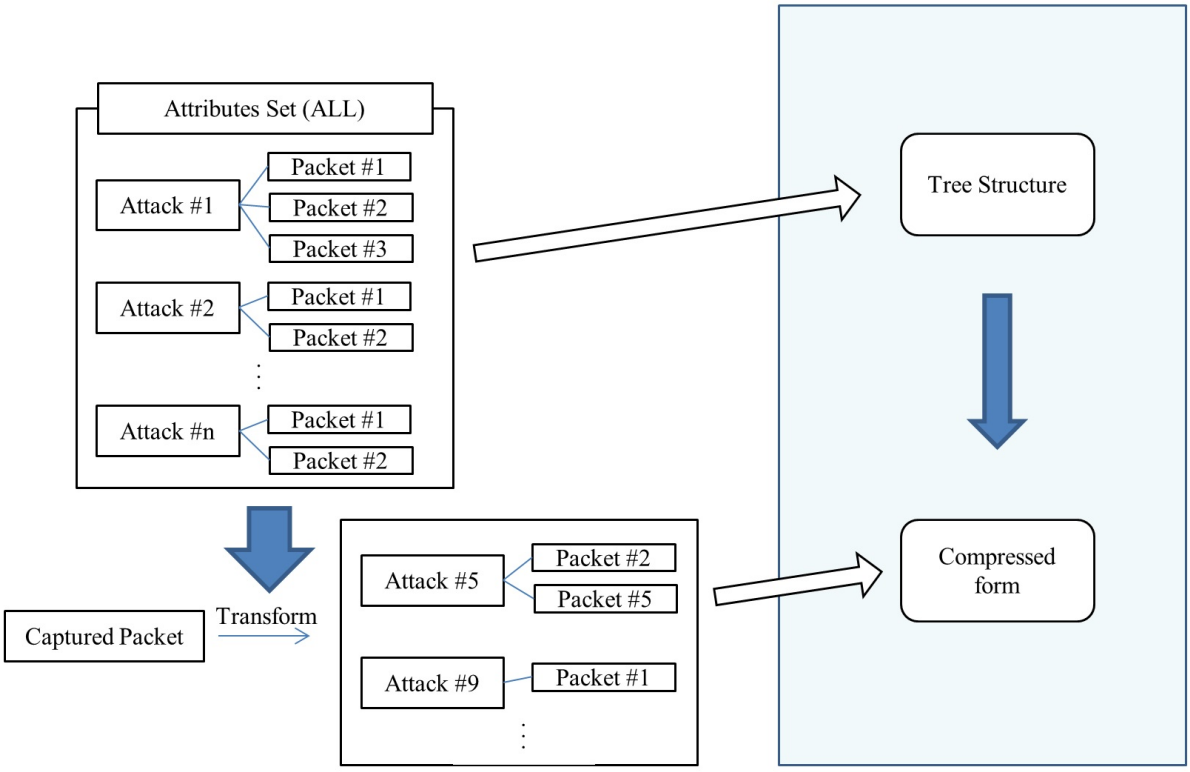

Figure 3: Packet collection method by attack type

For traceback in this case, it is necessary to know how many packets appropriate for attack types were generated among the classified packet information and in which path the generated attack types entered. In order to detect this, it is necessary to calculate the frequency of the data appropriate for the stored attack types. Search speed has a very important role because this process requires to read all the stored data within a certain time.

4.2 IP Traceback Analysis

Figure 4. shows the collected information by collection position as a table consisting of attack type, frequency of occurrence and occurrence path.

In this figure, we aims to perform traceback on attack type 7, which has the highest attack frequency in the destination. Paths are made by following the most frequent influx path in the same attack type. However, when there are several paths in the same attack type, this was made into a tree or graph or the highest path performs traceback. As shown in Figure 5, traceback is performed based on the analysis

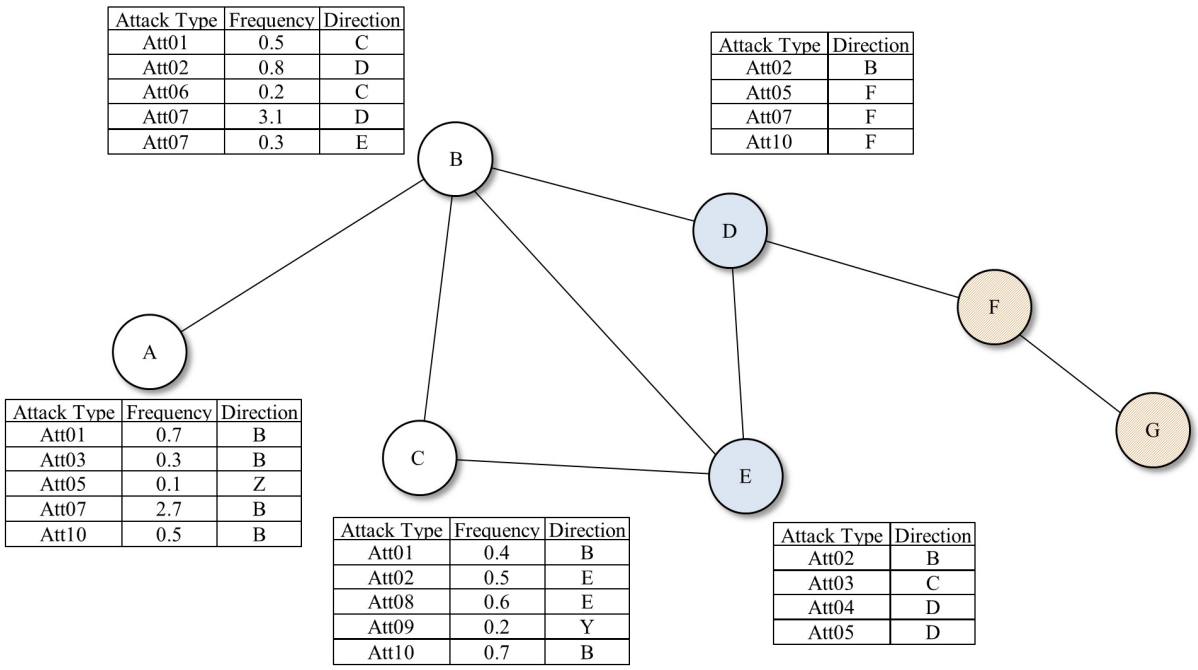

Figure 4: Attack type classification table by collection position

content in Figure 4 and was described as tree-shaped traceback graph.

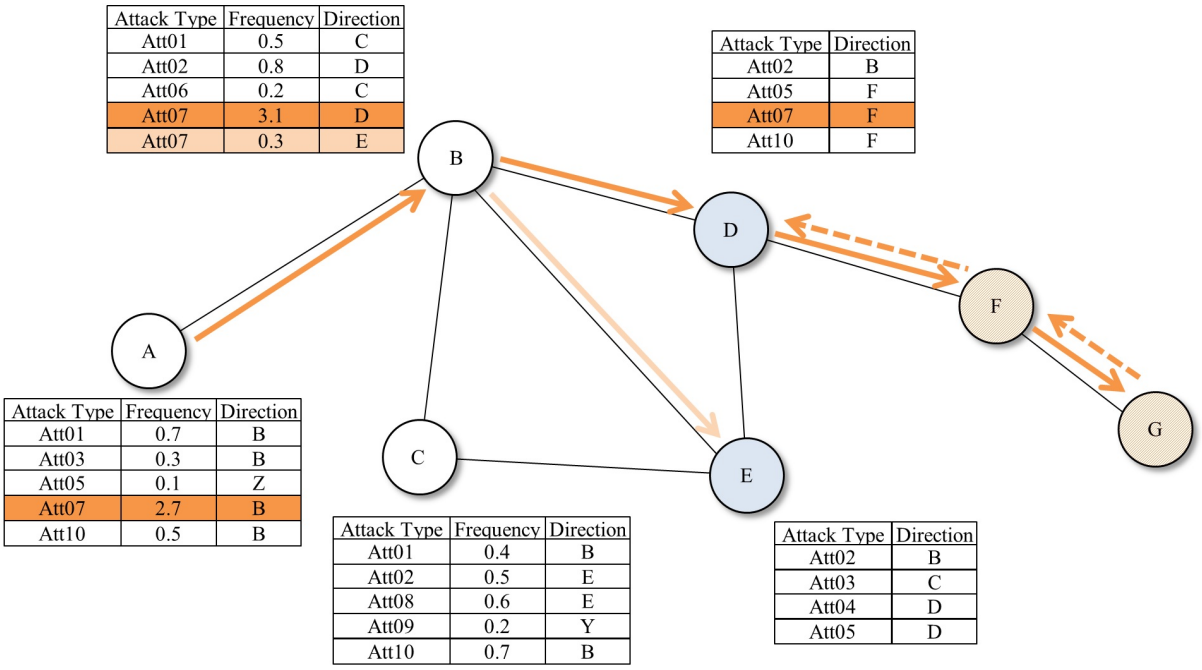

Figure 5: Traceback method using attack type classification table

Node D and node E in Figure 5 have no frequency of attack occurrence in the table. In such a case, it is not possible to collect entire information because partial information is missing or the system is

different. Node F and node G are in a state without any collected data at all. In such a case, several processes such as contacting the appropriate network manager, receiving traffic by additional collections and analyzing it are added.

4.3 Additional Collection

A traceback graph is generated by following the path with the actual traffic and additionally performing the collection and analysis. When the router without the table becomes the target of traceback, the collected traffic data is received by requesting a network manager. The received data was additionally analyzed for traceback. Although it is difficult to detect the accurate traceback point only by the information collected from the site of attack, it has the advantage of being able to traceback in comparison between the additionally collected data by SCS and their own information. Of course, not all networks have SCS. Therefore, it is possible to analyze after receiving the data from the manager who performs functions similar to SCS. Figure 6 shows there procedures.

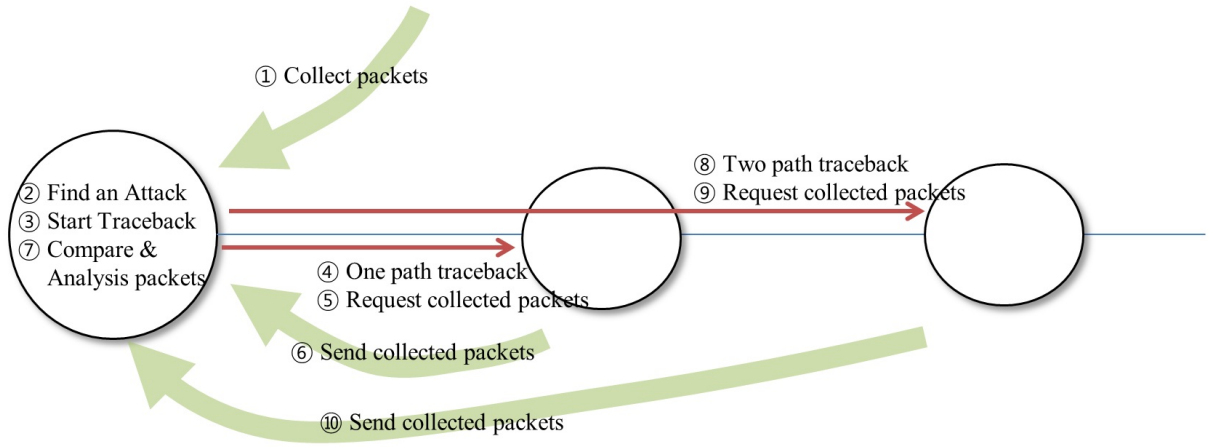

Figure 6: Additional data collection in position without collected data

The additionally received data can additionally be used for vulnerability analysis or rule update, and furthermore, the information collected from the external network can be continuously received by installing SSCS.

4.4 Advantages and Shortcomings

The traceback method proposed in this paper is one of the proactive methods to perform the traceback using data mining after the attacks in the framework with mass storage. It is a logging-based approach to be able to trace back not only DDoS attacks but also all kinds of attack packets among the proactive methods. The logging-based approach has the advantage of being able to perform an accurate traceback by analysis after the attacks and traceback all kinds of packets. In addition, it is possible to collect a large amount of data and traceback by reducing the large-scale cluster-based hierarchical structure and storage volume by attack type. Moreover, it can be applied not only to traceback but also only to rule making of other security devices and vulnerability analysis. Finally, this method proposed storage methods that was convenient for storage and search and reduced the disadvantage of the proactive method, an increase in the amount of calculation, by using data mining.

However, there has been not verification or simulation whether hierarchical cluster storage space can be overcome in increase in the data storage depending on the high-speed network, the disadvantage of

logging-based approach. In the future, this study will perform the calculation and simulation of this problem.

5 Conclusion

Methods to protect network attacks are to detect an attack and report it to the manager during it is made and traceback the source of the attack. The traceback is a way to search for sources of damage to the network or host computer. Traceback methods consist of reactive and proactive method. In these methods, proactive method consists of marking-based and logging-based approaches, and the logging method induces a serious storage overhead.

However, a system of capable of solving these problems through cluster-based mass storage, digestible packets and hierarchical collections was designed. It not only performs traceback but also communicate with analysis data of other security systems by using the logging-based methods. Thus, it can be used as a basic data to generate a new rule. It classifies attack types and stores only the analyzed information by using the framework capable of such mass storage. Such data mining can reduce an amount of calculation for storage volume and traceback.

5.1 Acknowledgments

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Plannig (2011-0029924)

References

[1] B. Al-duwairi and M. Govindarasu. Novel hybrid schemes employing packet marking and logging for ip traceback. *IEEE Transactions on Parallel and Distributed System*, 17(5):403–418, May 2006.

[2] Bahmani and Faezeh. A survey of interoperability in enterprise information security architecture frameworks. In *Proc. of the 2nd International Conference on Information Science and Engineering (ICISE'10), Hangzhou, China*, pages 1794–1797. IEEE, December 2010.

[3] N. Davuth and S.-R. Kim. Classification of malicious domain names using support vector machine and bi-gram method. *International Journal of Security and Its Aplications*, 7(1):51–58, January 2013.

[4] P. Do, H.-S. Kang, and S.-R. Kim. Improving a hierarchical pattern matching algorithm using cache-aware aho-corasick automata. In *Proc. of the 2012 ACM Research in Applied Computation Symposium (RACS'12), SanAntonio, Texas, USA*, pages 26–30. ACM, October 2012.

[5] C. Gong and K. Sarac. Toward a more practical marking scheme for ip traceback. In *Proc. of the 3rd International Conference on Broadband Communications, Networks and Systems (BROADNETS'06), San Jose, CA, USA*, pages 1–10. IEEE, October 2006.

[6] C. Gong and K. Sarac. A more practical approach for single-packet ip traceback using packet logging and marking. *IEEE Transactions on Parallel and Distributed System*, 19(10):1310–1324, October 2008.

[7] D. Gupta, D. S. Kohli, and R. Jindal. Taxonomy of tree based classification algorithm. In *Proc. of the 2nd International Conference on Computer and Communication Technology (ICCCT'11), Allahabad, India*, pages 33–40. IEEE, September 2011.

[8] T. Lee, W. Wu, , W, and Huang. Scalable packet digesting schemes for ip traceback. In *Proc. of the IEEE International Conference on Communications (ICC'04), Paris, France*, volume 2, pages 1008–1013. IEEE, June 2004.

[9] N. Lu, Y. Wang, F. Yang, and M. Xu. A novel approach for single-packet ip traceback based on routing path. In *Proc. of the 20th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP'12), Garching, Germany*, pages 253–260. IEEE, February 2012.

[10] S. Roy, A. Singh, and A. S. Sairam. IP Traceback in Star Colored Networks. In *Proc. of the 5th International Conference on Communication System and Network (COMSNET'13), Bangalore, India*, pages 1–9. IEEE, January 2013.

[11] A. Snoeren, C. Partridge, L. Sanchez, C. Jones, F. Tchakountio, B. Schwartz, S. Kent, , and W. Strayer. Single-Packet IP traceback. *IEEE/ACM Transaction on Networking*, 10(6):721–734, December 2002.

[12] Y. Wang, S. Su, Y. Yang, and J. Ren. A More Efficient Hybrid Approach for Single-Packet IP Traceback. In *Proc. of the 20th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP'12), Garching, Germany*, pages 275–282. IEEE, February 2012.

[13] T. Wong, K. Law, J. C. Lui, and M. Wong. An Efficient Distributed Algorithm to Identify and Traceback DDoS Traffic. *The Computer Journal*, 49(4):418–442, February 2006.

Author Biography

Ho-Seok Kang is a postdoctoral fellowship of the division of Internet and Multimedia Engineering at Konkuk University, Seoul, Korea. He received his Ph.D. degree in computer enginnering at Hongik University, Korea. His recent research interests are in network security, network protocol, mobile security, distributed algorithms and cloud computing.

Sung-Ryul Kim is a professor of the division of Internet and Multimedia Engineering at Konkuk University, Seoul, Korea. He received his Ph.D. degree in computer enginnering at Seoul National University, Korea. His recent research interests are in cryptographic algorithms, distributed algorithms, security in general, cloud computing, and data mining.