A Model for Determining the Origin of A Packet To Find Real Sources of Attacks

Challakonda Jaya Lakshmi Department of M.C.A TKR College Of Engineering &Technology, Meerpet, Hyderabad.

Aravind Puppala Department of M.C.A, TKR College Of Engineering & Technology, Meerpet, Hyderabad.

Prem chander Tudi Department of M.C.A, TKR College Of Engineering & Technology, Meerpet, Hyderabad.

Abstract

Internet Protocol (IP) trace back is the enabling technology to control Internet crime. In this paper, we present a novel and practical IP trace back system called Flexible Deterministic Packet Marking (FDPM) which provides a defense system with the ability to find out the real sources of attacking packets that traverse through the network. While a number of other trace back schemes exist, FDPM provides innovative features to trace the source of IP packets and can obtain better tracing capability than others. In particular, FDPM adopts a flexible mark length strategy to make it compatible to different network environments; it also adaptively changes its marking rate according to the load of the participating router by a flexible flow-based marking scheme. Evaluations on both simulation and real system implementation demonstrate that FDPM requires a moderately small number of packets to complete the trace back process; add little additional load to routers and can trace a large number of sources in one trace back process with low false positive rates. The built-in overload prevention mechanism makes this system capable of achieving a satisfactory trace back result even when the router is heavily loaded. The motivation of this trace back system is from DDoS defense. It has been used to not only trace DDoS attacking packets but also enhance filtering attacking traffic. It has a

wide array of applications for other security systems.

1. Introduction

Nowadays more and more critical infrastructures are increasingly reliant upon the Internet for operations. Attacks against Internet-connected systems are now so common place that Internet crime has become a ubiquitous phenomenon. Although a number of countermeasures and legislations against Internet crime have been proposed and developed, Internet crime is still on the rise. The dynamic, stateless, and anonymous nature of the Internet makes it extremely difficult to trace the sources of Internet crime, since the attacker can forge the source address field in an Internet Protocol (IP) packet. To find the real source of Internet attacks, we must possess the capability of discovering the origin of IP packets without relying on the source IP address field. This capability is called IP trace back. IP trace back systems provide a means to identify true sources of IP packets without relying on the source IP address field of the packet header, and are the major technique to find the real attack sources [1], [2].

In this paper, a novel and practical IP trace back system, Flexible Deterministic Packet Marking (FDPM), is presented. FDPM belongs to the packet marking family of IP trace back systems. The novel characteristics of FDPM are in its flexibility: first, it can adjust the length of marking field according to the network protocols deployed (flexible mark length strategy); second, it can also adaptively change its marking rate according to the load of the participating router by a flexible flowbased marking scheme. These two novel characteristics of FDPM make it more practical than other Current trace back systems in terms of compatibility and performance. simulation Both and real system implementation prove that FDPM can be used in real network environments to trace a large number of real sources, with low false positive rates, and with low resource requirement on routers.

2. Previous Work On IP Traceback

2.1. Problem Description

Let Ai, i $\in [0,n]$, be the attackers and V be the victim. The attackers and victim are linked by various routers Rj, j € [1, m]. The main objective of IP trace back problem is to identify the n routers directly connected to Ai. The key issue here is to completely identify the n routers with low false positive rates in a single trace back process (conducted by the same trace back point, e.g., V, for a certain period) because correlating the data in different trace back processes is not only extremely difficult but also meaningless for tracing a time-dependent event. In [3], it was stated that a practical IP trace back system should be able to identify a few hundred (10)sources/routers out of 1 million routers.

2.2. Current IP Trace back Schemes

Current IP trace back schemes can be classified into five categories: link testing, messaging, logging, packet marking, and hybrid schemes.

The main idea of the link testing scheme is to start from the victim to trace the attack to upstream links, and then determine which one carries the attack traffic [7], [8]. It consumes huge amount of resources, introduces additional traffic, and possibly causes denial of service when the number of sources needed to be traced increases. Messaging schemes use routers to send ICMP messages from the participating routers to destinations. For a high volume flow, the victim will eventually receive ICMP packets from all the routers along the path back to the source, revealing its location [9], [10], [11].

The disadvantages of messaging schemes are that the additional ICMP traffic would possibly be filtered by some routers, and huge numbers of packets are required by the victim to identify the sources. Logging schemes include probabilistic sampling and storing transformed information. Logging schemes maintain a database for all the traffic at every router within the domain and to query the database to identify the sources of an IP packet. Hash function or Bloom filter is used to reduce the data stored.

The main disadvantage of logging schemes is that they heavily overload the participating routers by requiring them to log information about every packet passing by, although it is claimed that it needs only a single packet to find its origin [12], [13], [14], [15].As this method overwrites some rarely used fields in IP header, it does not require modification of the current Internet infrastructure. This property makes it a promising trace back scheme to be part of DDoS defense systems[21].

2.3. Probablistic Packet Marking Schemes

Probabilistic Packet Marking (PPM) [16] is one stream of the packet marking methods. The assumption of PPM is that the attacking packets are much more frequent than the normal packets. It marks the packets with path information in a probabilistic manner and enables the victim to reconstruct the attack path by using the marked packets. PPM encodes the information in rarely used 16-bit Fragment ID field in the IP header. To reduce the data that is to be stored in 16 bits, the compressed edge fragment sampling algorithm is used.

Although PPM is simple and can support incremental deployment, it has many shortcomings that can seriously prevent it from being widely used. First, the path reconstruction process requires high computational work, especially when there are many sources. Second, when there are a large number of attack sources, the possible rebuilt path branches are actually useless to the victim because of the high false positives. Therefore, the routers that are far away from the victim have a very low chance of passing their identification to the victim because the information has been lost due to overwriting by the intermediate routers.

2.4. Deterministic Packet Marking schemes

Another stream of packet marking methods, which does not use the above probabilistic assumption and stores the source address in the marking field, is in the category known as the deterministic approaches, such as Deterministic Packet Marking (DPM) [26]Unlike PPM, deterministic approaches only keep the first ingress edge router's information in the marks.

This category of schemes has many advantages over others, including simple implementation, no additional bandwidth requirement, and less computation overhead. However, enough packets must be collected to reconstruct the attack path.

3. Flexible Deterministic Packet Marking Scheme

3.1. System Overview

The FDPM scheme utilizes various bits (called marks) in the IP header. The mark has flexible lengths depending on the network protocols used, which is called flexible mark length strategy. The source IP addresses are stored in the marking fields. The mark will not be overwritten by intermediate routers when the packet traverses the network.

Processing packets consume resources such as memory and CPU time of a participating router. Therefore, it is possible for a router to be overloaded when there are a large number of arrival packets waiting for FDPM to mark them.

The flexibility of FDPM is twofold. First, it can use flexible mark length according to the network protocols that are used in the network. Second, FDPM can adaptively adjust its marking process to obtain a flexible marking rate. This characteristic prevents a trace back router from the overload problems.

3.2. Utilization of IP Header

FDPM is based on IPv4. Possible IPv6 implementation of FDPM will involve adding an extension header in IPv6 packets, which is different with the IPv4 design.

Three fields in the IP header are used for marking; they are Type of Service(TOS), Fragment ID, and Reserved Flag. The TOS field is an 8-bit field that provides an indication of the abstract parameters of the quality of service desired. Fragment ID and Reserved Flag are also exploited. Fragment ID can be safely overloaded without causing serious compatibility problems As shown in Fig. 1, a total of 25(8 + 16 + 1) bits are available for the storage of

0	4 8 16 19						31
٧	Version IHL Type of Service			Total length			
	Identification				Flags	Fragment offset	
Γ	TTL Protocol			Header checksum			
	Source IP address						
	Destination IP address						
Options field (if any)							
	IP data						

Fig. 1. The IP header fields (darkened) utilized in FDPM.

mark information if the protected network allows overwriting on TOS. When considering the possibility that the TOS field may be unavailable partly or totally, the minimum number of the bits in the IP header is 16.

3.3. Encoding Scheme

Before the FDPM mark can be generated, the length of the mark must be determined based on the network protocols deployed within the network to be protected. According to different situations, the mark length could be 24 bits long at most, 19 bits at middle, and 16 bits at least. Therefore, the flexible length of the marks results in three variations of the Encoding scheme, which are named as FDPM-24, FDPM-19, and FDPM-16.

FDPM encoding scheme is shown in Fig. 2. The ingress IP address is divided into k segments and stored into k IP packets. The padding is used to divide the source IP address evenly into k parts. The segment number is used to arrange the address bits into a correct order.



3.4. Reconstruction Scheme

The reconstruction process includes two steps: mark recognition and address recovery. When each packet arrives at the point that requires reconstruction, it is first put into a cache. The cache can also output the packets to another processing unit, by this design the reconstruction methods can be applied in a parallel mode. This will be left as our future work.



Fig.4. FDPM reconstruction scheme.

The mark recognition step is the reverse process of the encoding process. If the RF is 0, the mark length is 24 (both TOS and ID are deployed). If the RF is 1, according to different protocols of TOS used, the mark length is 16 or 19. The second step, address recovery, analyzes the mark and stores it in a recovery table. It is a linked-list table; the number of rows is a variable, and the number of columns in the table is k, representing the number of segments used to carry the source address in the packets.

1.	Reconstruction at victim V, in network N
2.	for each coming packet p passing the reconstruction point
3.	mark recognition (length and fields)
4.	if all fields in one entry are filled
5.	output the source IP
6.	delete the entry
7.	else
8.	if same digest and segment number exist
9.	create new entry
10.	fill the address bits into entry
11.	else
12.	fill the address bits into entry



When the hash collision occurs, more than one entry may be created in order to keep as much information as possible. The advantage of this design is that it can reconstruct all possible sources but the disadvantage is it also brings possible irrelevant information.

3.5. Flow-Based Marking Scheme

The idea of flow-based marking is to selectively mark the packets according to the flow information when the router is under a high load. Therefore, it can reduce the packet marking load of a router but still maintain the marking and trace back function other applications, this overload prevention mechanism can be modified accordingly to target most possible attacking packets. The goal of flow-based marking is to mark the most possible DDoS attacking packets,.



The advantage of this is that it can be easily implemented in current router Architecture.

The simple data structures include a dynamic flow table T and a FIFO queue Q, as shown in Fig. 6. Each record in T stands for a flow. There are two load thresholds Lmax and Lmin for the trace back router. Lmax is the threshold that controls the whole packet marking process, which means the router will not mark any packet if its load exceeds this value. These two thresholds should be set according to Real situations in routers.

1.	if (load of router R > threshold Lmm)
2.	do not mark any packets
3.	turn on congestion control mechanisms
4.	else if (load of router R > threshold Lmin)
5.	turn on flow-based marking at R, edge interface A, in network N
6.	for each incoming packet p
7.	check npkts with same destination address of p from T
8.	if $(npkts == 0, means no such flow in T)$
9.	add a new entry in T , set its $npkts = 1$
10.	else
11.	npkts ++
12.	insert packet p into Q
13.	calculate marking probability pa
14.	with probability p_{ii} mark the packet (encoding procedure)
15.	if Q is full
16.	dequeue
17.	else
18.	mark all the packets at R, edge interface A, in network N

Fig. 7. Algorithm of FDPM flow-based marking scheme.

When flow-based marking is turned on, the probability of marking an incoming packet from a particular flow is roughly proportional to the flow's share of bandwidth through the router. We define this probability pa as

Pa =npkts- min(npktsi , i $\in \{1, n\}$)/ max (npktsi , i $\in \{1, n\}$)- min(npktsi

$$i \in \{1,n\}$$
) * Lmax _ L/Lmax _ Lmin
(1)

Where npkts is the number of packets in the flow containing current incoming packet, L is the current load of the router. This definition has Pa \cap [0, 1]. When the current load of the router L reaches Lmax, Pa becomes 0, which means no marking is performed.

Therefore, when calculating the marking probability Pa, we use the EWMA npkts which is defined as

npktsk =npktsk_1 + $(1 _\alpha)$ npktsk (2)

where α is the filter constant, which dictates the degree of filtering, In our experiments, this filter constant is set to 0.95.

TABLE 1

Relationship between the Parameters in FDPM and DPM

k		2	4	8	16	32
s		1	2	3	4	5
a		16	8	4	2	1
FDPM-	d	0	6	9	10	10
16	N_{max}	1	64	512	1024	1024
DBM	d	0	7	10	11	11
DPM	N_{max}	1	128	1024	2048	2048
FDPM-	d	2	9	12	13	13
19	N_{max}	4	512	4096	8192	8192
FDPM-	d	7	14	17	18	18
24	N _{max}	128	16384	131072	262144	262144

This value controls how many historical values of npkts are used.

4. Simulation: Trace Large-Scale Sources

4.1. Evaluation Measurement: Maximum Number of Sources

One goal of this research is to find the maximum number of sources that FDPM can trace in a single trace back process. This is a very important evaluation measurement when the trace back system is used to trace largescale sources. FDPM offers a stronger capability of tracing multiple attacker sources than other trace back schemes.

From this table, we can see under the ideal situation the maximum number of sources that can be traced in by FDPM is

262,144. Fig.8 shows a comparison of the maximum number of sources that can be traced under different encoding schemes by FDPM and by DPM.

4.2. False Positive Analysis

False positives of FDPM come from collision in hash functions, a situation that occurs when two distinct inputs into a hash function produce Identical outputs. If more than one edge



Fig. 8. Maximum number of sources that can be traced, if no hash collision exists.

router marks the IP packet with the same digest bits, then, i the victim end, the reconstruction will mix the marks from different routers and generate incorrect source IP addresses.

Let Z be the set of all integers. The domain of an IP address can be written as the set $U = \{x/x \in Z \land 0 \le 232\}$. The domain of a digest can be written as the set $W = \{x/x \in Z \land 0 \le x < 2d\}$, $d \in [0, 18]$. Mathematical expectation of different digests from different IP addresses can be written as

E[F] = 2d - 2d (1 - 1/2d)N (3)

where d is the digest bits, and N is the number of different IP addresses.Then, the expected number of different values in segment bits can be written as

 $E[S]=2a_2a (1 - 1/2a)Nd$, (4)

where a is the address bits. Then, the expected number of permutations that result in a given digest can be written as

 $\begin{array}{l} E[P]=(\ E[S]\)k/2d=(\ 2a\ 2a(\ 1\ -1/2a)Nc)k/2d \\ ; \qquad (5) \end{array}$

where k is the segment number. The number of valid reconstructed IP addresses has two parts, the ones recovered by different digests, and the ones kept by multiple entries with same digests by the aforementioned strategy. Therefore, the false positive rate can be written as

 $n = (N _ E[F]) E[P] - [N_E[F]Nd/2d + E[S]k/2a+d]/N$

 $= [N - 2d + 2d(1 - 1/2d)N][2a - 2a (1 _ 1/2a)Nd]k/2dN - [N - 2d + 2d(1 _ 1/2d)N]Nd/2dN - [2a - 2a(1 - 1/2a)Nd]k/2a + d N (6)$

4.3. Simulation Environment

An SSFNet simulator [42] is created to simulate the whole process of FDPM and gather experimental data for analysis. SSFNet is a collection of Java components used for modeling and simulation of IPs and networks. An experimental network topology is set up according to a real network as it is shown in Fig. 9. The simulated FDPM system is installed on all the routers in the network. Three new Java packages are embedded into the SSFNet simulator, which are the Encoding subsystem, the Reconstruction



Fig. 9. Network topology in simulation.

In the Encoding subsystem, the hash function must be chosen carefully because hash collision is one of the main factors affecting the trace back performance in terms of the maximum number of sources that can be traced.

Three general-purpose hash functions, the MD5 hash function [44], the PJW hash function [45], and the BKDR hash function [46], are selected to test the effectiveness of hashing in FDPM. We chose these functions because they can be implemented in any programming language and are fast with good distribution capability.

4.4. Collision in Hash Functions

We define noncollision rate λ as the percentage of the nonrepeated hashed values in the total hashed values. Fig. 10 shows the average noncollision rate of the hashed digest in the trace back experiments. According to the above test on collision in hash functions, we can further obtain the average maximum number of sources that FDPM can actually trace. This is an important feature of FDPM of being a practical trace back system because, to our knowledge, no existing system can trace



percent.

such a large number of sources in a single trace back process. In this case, FDPM-19 can trace1,495 sources when k=4. However, the maximum number of sources that DPM can trace is 49 when k = 4.

From Fig. 11, we can also derive the optimal segment number k to achieve the maximum number of sources that can be traced.Collision in hash functions plays an important role in improving the maximum number of sources that can be traced.

5. Simulation: Overload Prevention

5.1.Evaluation Measurements: Marked Rate and Number of Packets Needed to Trace One Source

The overload prevention mechanism is important to all packet marking trace back schemes. The evaluation measurements of rating the effectiveness of the flow-based marking scheme are the marked rate β , and the number of packets needed to trace one source NN. Theoretically, if all incoming packets are marked, and there is no hash collision problem, then the expected number of packets needed to trace one source can be a Coupon Collector problem [41] decided by the number of segment used k, as

 $E[NN] = K(1/k + 1/k - 1 + \dots + 1).$



Fig. 12. The relationship between the number of packets needed to trace one source NN and the marked rate β for the flow-based marking scheme and the random marking scheme in simulation. (a) k = 2, γ = 0.1. (b) k= 8, γ = 0:5.

5.2.Flow-Based Marking versus Random Marking

When the load of a router exceeds a certain threshold, the router has to reduce the marking rate in order to alleviate the load. If the packets are marked in a random manner the Victim which possesses reconstruction will use more packets to reconstruct the sources than the flow-based marking scheme.

Fig. 12a shows the number of packets needed to trace one source NN and the marked rate β of all the packets passing through the router in the flow-based marking scheme and random marking scheme. Fig. 12b shows the marking efficiency in the flow-based marking scheme when the router uses eight packets to carry a source IP address (k = 8) and the percentage of attacking packets γ = 0.5. The expected number of packets needed to trace one source E[NN] is 2 when k = 8.

From Fig. 12, we find that when the router has to reduce its load of packet marking, the flowbased marking scheme performs much better than the random marking scheme in terms of the number of packets needed to trace one source NN and the marked rate β .

The generic relationship between NN and E[NN] can be written as

NN =
$$\alpha$$
 (β , γ , λ)E[NN] (8)

Where a is a function of β , γ and λ .

5.3. Percentage of Attacking Packets

Fig. 13 shows the relationships between the marked rate β , the number of packets needed to trace one source NN and the percentage of attacking packets γ . First, from the figures, we can see that as a higher percentage of attacking packets lead more packets to be marked, less packets are needed



Fig. 13. Relationships between the marked rate β , the number of packets needed to trace one source NN, and the percentage of attacking packets γ , in simulation. (a) k = 2. (b) k = 8.

6. Real System Implementation

6.1.Evaluationasurements: Number of Packets Needed to Trace One Source and Maximum Forwarding Rate

Currently, most existing works on IP trace back are based on simulation or thoretical analysis. It is very difficult to test the real performance of a trace back scheme if only simulation is conducted. The main evaluation measurements we used are the marked rate β , the number of packets needed to trace one source NN, and the maximum forwarding rate θ max.

at the reconstruction end when the percentage



Fig. 14. The relationship between the number of packets needed to trace one source NN and the marked rate β for the flow-based marking scheme and the random marking scheme in real system implementation. (a) k = 2, γ = 0:1. (b) k = 8, γ , γ = 0.5.

We used the Click modular router [47] to implement our FDPM on PC-based router. Click router is a software architecture running on PCs for building flexible and configurable routers, which is assembled from packet processing modules called elements.

6.2. Number of Packets for Reconstruction

Fig. 14 shows the relationship between the number of packets needed to trace one source NN and the marked β for flowbased marking scheme and random marking scheme in Click router implementation. The condition of Fig. 14a is that the router uses two packets to carry a source IP address (k = 2) and the percentage of attacking packets $\gamma = 0.1$. The condition of Fig. 14b is that the router uses eight packets to carry a source IP address (k = 8) and the percentage of attacking packets $\gamma = 0.5$.



Fig. 15. Maximum forwarding rate of the Click router.

6.3. Maximum Forwarding Rate

This section evaluates FDPM-enabled router's performance of forwarding IP packets under different conditions. Fig. 15 shows the maximum forwarding rate θ max for the raw Click router without any packet marking function. Fig.16 shows when k= 8, the curve of maximum forwarding rate θ max of an FDPM enabled router and the curve when all the packets are marked, which is defined as the all marking scheme.





Table 2 shows the relationship between the percentage of attacking packets γ and the maximum forwarding rate θ max of both an FDPM-enabled router and all marking scheme.

TABLE 2

Relationship between the Percentage of Attacking Packets and the Maximum Forwarding Rate

Percentage of at- tacking packets γ	$ heta_{max}$ of flow- based marking scheme	$ heta_{max}$ of all marking scheme		
1	65412	58423		
0.9	66144	59104		
0.8	65252	57451		
0.7	64099	56482		
0.6	65186	57412		
0.5	64230	54132		
0.4	63701	55265		
0.3	63383	52102		
0.2	64163	57412		
0.1	67170	56325		



Fig. 17. Relationships between the marked rate β , the number of packets needed to trace one source NN, and the percentage of attacking packets γ in real system implementation. (a) k= 2. (b) k = 8.

6.4. Percentage of Attacking Packets

Fig.17 shows in real system implementation the relationships between the marked rate β , the number of packets needed to trace one source NN, and the percentage of attacking packets γ . The performance of FDPM in real system implementation is slightly better than in simulation.

7. Conclusion

FDPM is suitable for not only tracing sources of DDoS attacks but also DDoS

detection. The main characteristic of DDoS is to use multiple attacking sources to attack a single victim.

In FDPM, the marks in packets do not increase their size; therefore, no additional bandwidth is consumed. Moreover, with the overload prevention capability, FDPM can maintain the trace back process when the router is heavily loaded, where as most current trace back schemes do not have this overload prevention capability. FDPM requires little computing power and adaptively keeps the load of routers in a low degree. Where compatibility is concerned,

Compared with other IP trace back schemes, FDPM provides more flexible features to trace IP packets than other packet marking schemes, and can obtain better tracing capacity.

To summarize this paper, we list our major contributions here:

1. A novel and practical packet marking trace back system, incorporating a flexible mark length strategy and flexible flow-based marking scheme, is proposed.

2.Simulation and real system implementation show FDPM produces better performance than any other current trace back scheme in terms of false positive rates, the number of packets needed to reconstruct one source, the maximum number of sources that can be traced in one trace back process, and the maximum forwarding rate of trace backenabled routers.

Acknowledgments

The Successful Completion of any task would be incomplete without expression of simple gratitude to the people who encouraged our work. Though words are not enough to express the sense of gratitude towards everyone who directly or indirectly helped in this task.

I thankful to this Organization TKR College of Engineering & Technology, which provided good facilities to accomplish my work and would like to sincerely thank to our Principal for giving great support, valuable suggestions and guidance in every aspect of my work.

REFERENCES

[1] H. Farhat, "Protecting TCP Services from Denial of Service

Attacks," Proc. ACM SIGCOMM Workshop Large-Scale Attack

Defense (LSAD '06), pp. 155-160, 2006.

[2] H. Wang, C. Jin, and K.G. Shin, "Defense against Spoofed IP Traffic Using Hop-Count Filtering," IEEE/ACM Trans. Networking, vol. 15, no. 1, pp. 40-53, 2007. [3] M.T. Goodrich, "Efficient Packet Marking for Large-Scale IP Traceback," Proc. Ninth ACM Conf. Computer and Comm. Security (CCS '02), pp. 117-126, 2002. [4] H. Aljifri, "IP Traceback: A New Denialof-Service Deterrent," IEEE Security and Privacy, vol. 1, no. 3, pp. 24-31, 2003. [5] A. Belenky and N. Ansari, "On IP Traceback," IEEE Comm., vol. 41, no. 7, pp. 142-153, 2003. [6] Z. Gao and N. Ansari, "Tracing Cyber Attacks from the Practical Perspective," IEEE Comm., vol. 43, no. 5, pp. 123-131, 2005. [7] H. Burch and B. Cheswick, "Tracing Anonymous Packets to Their Approximate Source," Proc. 14th Systems Administration Conf. (LISA '00), pp. 319-327, 2000. [8] R. Stone, "CenterTrack: An IP Overlay Network for Tracking DoS Floods," Proc. Ninth USENIX Security Symp. (Security '00), pp. 199-212, 2000.