available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

# Extensions to the source path isolation engine for precise and efficient log-based IP traceback

Egon Hilgenstieler [a], Elias P. Duarte, Jr. [a,*], Glenn Mansfield-Keeni [b,1], Norio Shiratori [c,2]

[a] Federal University of Paraná, Dept. Informatics, P.O. Box 19018, Curitiba 81531-980, PR Brazil
[b] Cyber Solutions Inc., ICR Bldg, 6-6-3, Minami Yoshinari, Aoba-ku, Sendai, 989-3204 Japan
[c] Tohoku University – RIEC, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

ARTICLE INFO

*Article history:*
Received 18 September 2009
Received in revised form
24 November 2009
Accepted 28 December 2009

*Keywords:*
Denial of service
Packet tracing
Traceback
Attack Response
Traffic logging

ABSTRACT

IP traceback is used to determine the source and path traversed by a packet received from the Internet. In this work we first show that the Source Path Isolation Engine (SPIE), a classical log-based IP traceback system, can return misleading attack graphs in some particular situations, which may even make it impossible to determine the real attacker. We show that by unmasking the TTL field SPIE returns a correct attack graph that precisely identifies the route traversed by a given packet allowing the correct identification of the attacker. Nevertheless, an unmasked TTL poses new challenges in order to preserve the confidentiality of the communication among the system's components. We solve this problem presenting two distributed algorithms for searching across the network overlay formed by the packet log bases. Two other extensions to SPIE are proposed that improve the efficiency of source discovery: separate logs are kept for each router interface improving the distributed search procedure; an efficient dynamic log paging strategy is employed, which is based on the actual capacity factor instead of the fixed time interval originally employed by SPIE. The system was implemented and experimental results are presented.



## 1. Introduction

Several attacks are reported daily in the Internet and several tools are widely available for attackers to disable network services either by exploiting software or hardware implementation bugs, or simply by flooding the network with legitimate requests. Denial of Service (DoS) attacks are frequently reported (SecurityStats.com, 2009), in which the network and/or a server is overwhelmed with heavy traffic. Other types of attacks, however, can be orchestrated using a significantly smaller amount of packets. There are attacks that can disable a network service with just a single packet (Microsoft Corporation, 2008). After an attack is detected and an attack packet is isolated, the next step is to determine its source. Unfortunately, it is not possible to reliably determine the source of a received IP packet, as the protocol does not provide authentication of the packet based on the source address field, which can be easily faked (IP Spoofing). Furthermore the Internet routing infrastructure also does not keep information about forwarded packets.

IP traceback was proposed in order to allow the discovery of the source of a packet received from an IP network (Belenky and Ansari, 2003; Gao and Ansari, 2005). Several approaches for IP traceback have been proposed, two of which have been considered the most important: Probabilistic Packet Marking (PPM), and log-based traceback. In PPM (Dean et al., 2002;

* *Corresponding author*. Tel.: +55 41 3361 3656; fax: +55 41 3361 3205.
E-mail address: elias@inf.ufpr.br (E.P. Duarte Jr.).
[1] Tel.: +81 22 303 4012; fax: +81 22 303 4015.
[2] Tel.: +81 22 217 5420; fax: +81 22 217 5426.

0167-4048/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cose.2009.12.011

Goodrich, 2002, 2008; Savage et al., 2000; Yaar et al., 2005; Gao and Ansari, 2007; Belenky and Ansari, 2007; Wong et al., 2008), routers mark packets chosen with a predefined probability with path information, typically employing header fields that are seldom used. This approach presents several challenges, such as the complexity of path reconstruction and the convergence of the traceback operation.

In log-based traceback (Keeni, 2009; Hazeyama and Kadobayashi, 2003; Sung et al., 2008) packet logs are kept throughout the network, ideally one per segment. The SPIE architecture (Source Path Isolation Engine) (Snoeren et al., 2001, 2002) is a log-based traceback that allows the path of a packet to be traced. The logs are not kept by the routers themselves, but by a packet monitor that listens to a router interface. The set of packet monitors form an overlay network that allows the source of individual IP packets to be determined. Hybrid approaches that are at the same time log-based and employ PPM have also been presented (Cong and Sarac, 2005).

The general goal of log-based traceback is to build an *attack graph*, given an IP packet, its approximate time of receipt and its destination, which is usually called the *victim*. The attack graph consists of vertices that represent nodes (routers and hosts) that have processed the packet, and the links through which the packet was transmitted. *False positives* are the nodes of the attack graph that have not really processed the packet. False positives can occur, for example, if a router is subverted by an attacker.

Several problems arise when defining a traceback architecture. The packets can be modified during the routing process; some possible transformations are described in RFC 1812 (Baker, 1995) such as packet fragmentation, options processing, ICMP (Internet Control Message Protocol) packet processing and packet duplication. Privacy issues are also important in the project of a traceback architecture. The packet's content should be properly protected. Furthermore, as a traceback architecture possibly requires the cooperation of several autonomous system (AS) it is desirable that even the packet metadata should be protected.

In this paper we present an approach for improving the precision and efficiency of SPIE. The proposed approach allows SPIE to return an attack graph that precisely identifies the route traversed by a given packet allowing the correct identification of the attacker. We show that without the extensions SPIE can return misleading attack graphs in some particular situations, which may even make it impossible to determine the real attacker. We show that by unmasking the TTL field SPIE returns a correct attack graph that precisely identifies the route traversed by a given packet allowing the correct identification of the attacker. Nevertheless, an unmasked TTL poses new challenges in order to preserve the confidentiality of the communication among the system's components. We present two new traceback algorithms which guarantee that communication among the system's components preserves the confidentiality of the packet's information.

Two other features are proposed to improve the efficiency of the traceback process: separate logs are kept for each router interface improving the distributed search procedure, having a strong impact on the cost of the traceback operation, as the number of requests is reduced to the minimum. Finally, an efficient dynamic log paging strategy is proposed, which is based on the actual capacity factor instead of the fixed time interval originally employed by SPIE.

The traceback system was implemented and experimental results are presented. Three metrics are evaluated: the precision of the obtained attack graph, that allows the correct determination of not only the packet source but also the entire route traversed by the packet; the cost of the traceback operation, measured by the number of requests in the network, and the time-frame from which a received packet can still be traced.

The contributions of this paper can be thus summarized as:

- We show that SPIE can return misleading attack graphs in some particular situations, which may even make it impossible to determine the real attacker.
- We show a simple solution to the problem (unmasking the TTL field) which nevertheless poses new challenges in order to preserve the confidentiality of the communication among the system's components.
- Two new traceback algorithms are proposed which guarantee that communication among the system's components preserves the confidentiality of the packet's information.
- We show that keeping separate logs for each router interface reduces the number of requests to a minimum.
- A dynamic and efficient approach for paging logs to secondary memory is proposed based on the log's actual capacity factor, instead of a fixed time interval.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 gives an overview of SPIE. In section 4 we describe the situation that causes SPIE to return incorrect attack graphs; the solution to the problem is given in section 5. Section 6 is devoted to the extensions proposed to improve SPIE's efficiency. The implementation as well as experimental results are presented in section 7. Conclusions follow in section 7.

## 2. Related work

The earlier approaches for tracing packets in the Internet aimed at determining the route of a packet stream rather than of a single packet. In Burch (2000) some networks are systematically flooded with packets. In this way it is possible to observe variations on the received packet stream and infer the traversed route. In other early efforts, the routers provide partial information about the route traversed to the end-hosts, employing a subset of the packets of a given flow. The end host can then reconstruct the packet path after receiving a large enough amount of packets.

Those early approaches required large amounts of packets in order to infer the traversed route. Those proposals cannot be applied to discover the source of attacks conducted with a smaller amount of packets. Later, other approaches for tracing back individual IP packets were developed.

A strategy based on the Internet Control Message Protocol (ICMP) has been proposed (Wu et al., 2001), in which a new ICMP message is defined that is sent randomly by routers along a path traversed by a given packet, either to the destination or origin of that packet. The communication overhead of this approach has been considered an important issue.