基于 AMS 反向追踪的二维门限重构算法

杨雪芹，裴昌幸，朱畅华，陈 南，易运晖
(西安电子科技大学 综合业务网理论与关键技术国家重点实验室，陕西 西安 710071)

摘要：提出了基于 AMS(Advanced Marking Schemes)的二维门限重构算法. 该算法在重构过程中设置了二维门限 $m_{f,d}$，通过判断攻击包的边域和节点的 Hash 值匹配情况，快速而准确地重构出攻击路径，从而缩短了整体重构时间和业务开销，并使准确度和稳定性得以提高，与传统方法相比，该二维门限重构算法使得反向追踪性能明显改善.

关键词：AMS 算法；反向追踪；快速重构
中图分类号：TP393 **文献标识码**：A **文章编号**：1001-2400(2006)02-0304-03

AMS based reconstruction algorithm with the two-dimentional threshold for IP traceback

YANG Xue-qin, *PEI Chang-xing*, *ZHU Chang-hua*, *CHEN Nan*, *YI Yun-hui*
(State Key Lab. of Integarted Service Networks, Xidian Univ., Xi'an 710071, China)

Abstract: We present a new reconstruction algorithm based on the Advanced Marking Scheme(AMS). It works with a two-dimentional threshold $m_{f,d}$, decides if a node is on the attack path by judging the situation of the edge of the packet and the Hash value match, then consequently reduces the time of reconstruction and overhead and improves accuracy and stability. Compared with the conventional methods, two-dimentional threshold reconstruction algorithm impves the performance of the IP traceback technique obviously.

Key Words: advanced marking scheme; IP traceback; reconstruct rapidly

在各种针对网络的安全威胁中，拒绝服务攻击(Denial of Service，简称 DoS)因危害巨大、难以防御等特点，成为黑客经常采用的攻击手段. IP 追踪主要针对 DoS 攻击. 这种攻击不需要目标主机回复数据，因此攻击主机很可能伪装源 IP 地址，隐藏其真实来源. 这也是 IP 追踪的难点所在. 尽管 IP 包头中的地址是虚假的，每个 IP 包仍然要经过从攻击方到受害方之间的路由器来转发，找到并记录下这些转发路由器，就可恢复出攻击所经过的路径，这是反向追踪的基本思路.

现有的追踪方法主要有 ICMP Traceback，Overlay Network，Hash-based IP Traceback，Controlled flooding，Traceback with IPSec，Probabilitic Packet Marking 等. 文献[1]对这些方法进行了分析比较，发现从人工管理量[2]、网络负载[3]、路由器负载、对现有协议的要求等多个方面来看，概率数据包标记算法是性能比较好的一种反向追踪方法. 数据包标记算法经过了从节点增加算法、节点采样算法、边界采样算法的发展，2001 年 Stefen Savage 提出了压缩边界分片采样算法(Fragment Marking Scheme，简称 FMS)[4]，解决了 64 位边标记信息的问题. Dawn Xiaodong Song 和 Adrian Perrig 对 FMS 算法进行改进，提出了 AMS(Advanced Marking Scheme)算法[5]，解决了 FMS 算法高计算负载和高误报率的问题.

笔者提出的二维门限重构算法，在 AMS 算法的基础上进一步提高重构过程的效率，并进一步减少了重构过程的开销，缩短了重构时间.

收稿日期：2005-09-30
基金项目：国家自然科学基金重点项目(60132030)；西安市工业科技攻关项目(GG04018)
作者简介：杨雪芹(1979-)，女，西安电子科技大学硕士研究生.

1　基于 AMS 反向追踪的二维门限重构算法

AMS 算法是一种边标记算法，通过存储每个 IP 地址的 Hash 值而不是 IP 地址本身进一步减小存储空间需求. 算法假定目标主机有所有路由器的完整网络图，在边分片重组后，通过比较结果 IP 地址的 Hash 值和网络图上路由器的 IP 地址的 Hash 值，进行攻击路径重构. 二维门限重构算法是在 AMS 算法的重构过程中实施的. 重构过程中采用了一个二维门限 $m_{f,d}$ 方案，只有攻击包中 $m_{f,d}$ 个包的 Hash 值和图表中节点 u 的 Hash 值匹配时，u 才被加入到攻击路径中.

1.1　标记过程

文献[6]指出，在 IP 包头的格式中，16 bit的分片标识位很少被使用(只有 0.25% 的包被分片)，因此数据包标记算法考虑用其存放标记信息. 将 16 bit中 5 bit作为 distance 域用于标记路径长度(大多数情况 $2^5=32$ 跳已经是可能的最大跳数了). w 比特作为 FlagID 域用来标记所选的 Hash 函数，为了将候选的边界信息，即两个路由器的 32 位 IP 地址放入 $11-w$ 位的字段中，可利用 Hash 函数将 IP 地址压缩到 $11-w$ 位，进而利用 XOR 运算，将两个 $11-w$ 位的地址信息放入 IP 包头中. 为了减小碰撞，降低误警率，选取多个 Hash 函数(图中选取 Hash 函数的数目为 2^w)，IP 头的编码如图 1 所示.

Distance (5 bit)	FlagID (w bit)	Edge (11-w) bit

图 1　IP 头的编码

当路由器 R_i 打算标识数据包时，选择一个 w bit 的随机数 x 写入标记域(FlagID)，利用 $g(<x,R_i>)$ 作为其 IP 地址的编码写入边(Edge)域. 每个路由器以概率 q 标记前向数据包. 如果路由器 R_i 打算标识数据包 P，就将 $g(<x,R_i>)$ 写入边(Edge)域，0 写入数据包 P 的距离(Distance)域. 否则，若距离域已经为 0(表明前面的路由器已经标识数据包)，路由器将生成 w bit 的随机数 y 写入标记域，将不同的 Hash 函数生成的 IP 地址编码 $g'(<y,R_i>)$ 的值和原来边域中的值进行异或，用异或的结果改写边域. 如果路由器不打算标识数据包，就总是增加距离域的值. 相邻两个路由器的异或是对目标主机的上游路由器图表中两个路由器之间边的编码.

由于 $a\oplus b\oplus a=b$，可从距离目标主机一跳的路由器开始，逐跳地对以前经过的路由器进行解码.

标记算法的伪代码如下：

```
Marking procedure at route R_i
for each packet P
let u be a random number from [0,1]
if ( u ≤ q) then
    let x be a random number from [0,7)
    P. fID←x
    P. distance←0
    P. edge←g( <l,R_i> )
else
    if(P. distance = =0) then
    P. edge←P. edge⊕g'( <P. fID, R_i> )
    P. distance←P. distance + 1
```

1.2　重构过程

攻击路径的重构过程是在受害主机处完成的. 受害主机利用上游路由器图表作为路线图，从 root 开始执行宽度优先搜索. 将 distance 为 d，FlagID 为 l ($l\in 0,2^w-1$) 的边片段表示为 $\psi_{d,l}$. 受害主机列出路线图中所有距其一跳的路由器 R_i，对于 l 的取值 l ($l\in 0,2^w-1$)，检查哪个路由器 IP 地址的 Hash 值 $g(<x,R_i>)$ 和边域集合 $\psi_{0,1}$ 中的值匹配，将匹配的 IP 地址集合表示成 S_0. S_d 表示重构出的图中距离受害主机 d 跳的路由器的 IP 地址的集合. 对于 $\psi_{d+1,l}$ 中的每条边 x 和 S_d 中的每个元素 y，受害主机计算 $z=x\oplus g(<l,y>)$. 对于 l 的所有可能取值，受害主机检验图表 G_m 中 y 的孩子 R_j 的 IP 地址的 Hash 值，$g(<l,R_j>)$ 是否与 z 相等. 如果目标主机找到一个匹配的 IP 地址 R_u，就将 R_u 加到集合 S_{d+1} 中去. 受害主机重复该过程，直到达到标识包中的最大距离max d，完成路径的重构.

重构算法的伪代码如下：

```
Reconstruction procedure at victim v ：
  let S_d be empty for 0 ≤ d ≤ max d
  for ecah chile R of v in G_m
  let count = 0
  for l : = 0 to 2^w - 1
    if g( < l,R > ) ∈ ψ_{0,1} then
      count = count + 1
```

```
    if count > m_{f,0} then
      insert R into S_0
  for d := 0 to max d - 1
    for each y in S_d
      for each child u for y in G_m
        let count = 0
        for l := 0 to 2^w - 1
          for each x in ψ_{d+1,l}
            z = x ⊕ g'(< l,y >)
          if g(< l,u >) = z then
            count = count + 1 ; break
    if count > m_{f,d} then
      insert u into S_{d+1}
  output S_d for 0 ≤ d ≤ max d
```

2 算法的性能分析

对 AMS 算法进行分析,假设路由器的标记概率是 p,一个数据包被和受害主机距离为 d 和距离为 $d-1$ 的两个相邻路由器标记(数据包中标记的是和受害主机距离为 d 和距离为 $d-1$ 的两个相邻路由器组成的边,即和受害主机距离为 $d-1$ 的一条边信息) 而不被其下游路由器更改的概率为 $p(1-p)^{d-1}$,当通过这条边的数据包数目为 N 时,受害主机收到的标有这条边路径信息的数据包数目的数学期望为

$$M = N p(1-p)^{d-1} \quad . \tag{1}$$

由此可见受害主机收到标有与其距离为 $d-1$ 的边的信息的数据包数目的数学期望值 M 和通过该边的数据包数目 N,路由器的标记概率 p 和距离 d 有关.

对于式(1)对标记概率 p 求导,当 $p = 1/d$ 时,M 的值最大,通常路由器的标记概率 p 取 $d = 25$ 时的经验值 0.04.

通过对式(1)的分析可看到,对于和受害主机距离为 $d-1$ 的边,通过的数据包数目 N 越小,该边路径上流量的百分比越小,受害主机收到的标有该边信息的采样数据包数目越小;d 值越大,即标记边和受害主机之间的距离越远,受害主机收到的标有该边信息的采样数据包数目越小.

可见在路径重构时采用相同的门限 m 是不公平的. 根据上面的分析考虑应用二维可变门限 $m_{f,d}$. 对于和受害主机距离相等的多条边,哪条边上攻击流量占总攻击流量的百分比越小,重构时该边门限数据包的数目 $m_{f,d}$ 越小;距离受害主机越远的边,重构时门限数据包数目 $m_{f,d}$ 越小. 这样,对于距离受害主机较远的边,或者是攻击流量很小的边,路径重构时门限数据包的数目较小,即对于这样的边采样数据包,受害主机收到数目较少时,就可认为该边在攻击路径上,从而提高路径重构的效率和准确性.

3 仿真结果分析

仿真是在 VC 平台上用 C++ 语言编写的程序进行的. 仿真过程中,w 的值取为 3,标记概率 q 取为经验值 0.04. 在路由器处对包进行标记,在受害主机处利用标记的包进行路径重构.

首先对两种情况下重构长度不同的路径所需要的开销进行了仿真对比,用一个攻击者在距离受害主机不同距离的情况下对受害主机发起攻击,然后测试重构此路径需要的攻击包的数量,正确率在 95% 以上. 结果如图 2 所示,从图中能够看出相同的标记情况下重构不同长度的攻击路径,二维门限算法所需的开销远远小于 AMS 算法的开销. 之所以会出现这种情况,是因为 AMS 中 m 取的是定值. 由前面的分析可知距离受害者较近的路由器被标记的概率大于距离受害者较远的路由器,所以如果 m 的取值能够重构出距离受害者较近的路由器,那么距离受害者较远的路由器的门限 m 就不用那么大. 在二维门限算法中 $m_{f,d}$ 的取值可参考 AMS 算法中 m 的取值来确定最佳门限.

得到两种开销的对比结果后,又对两种情况下重构路径的误判率进行了对比. 选择的网络拓扑含有距离受害者的距离为 4~29 跳,每一跳都有 10 个攻击者,发包的数据能够使其重构路径的正确率保证在 95% 以上. 结果如图 3 所示,从图中可看出相同情况下二维门限算法错误的重构路径小于 AMS 算法,在重构过程中发现 AMS 算法的稳定性比二维门限算法要差,分析可知 m 取定值则灵活性很差,它不能根据链路的流量以及路由器距离受害者的距离来调节门限值,当流量或者距离发生变化时,它不能适应这种变化,就不稳定,因

(下转第 321 页)