软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(6):1334−1345 [doi: 10.3724/SP.J.1001.2013.04279]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于加权累积和检验的加密流量盲识别算法*

赵 博, 郭 虹, 刘勤让, 邬江兴

(国家数字交换系统工程技术研究中心,河南 郑州 450002)

通讯作者: 赵博, E-mail: chaoxbang@gmail.com

摘 要: 针对加密流量的在线普适识别问题,提出一种基于加权累积和检验的时延自适应加密流量盲识别算法.利用加密数据的随机性特点,对网络报文逐一实施累积和检验,根据报文长度将结果进行加权综合.无需解密操作,也无需匹配特定内容,实现了对加密流量的普适识别.可动态调整报文的检测数量,以达到时延和准确率的统一,实现在线识别.仿真结果显示,对公开和未公开的加密协议流量,识别率均可达到 90%以上.

关键词: 流量分类;加密流量识别;累积和检验

中图法分类号: TP393 **文献标识码**: A

中文引用格式: 赵博,郭虹,刘勤让,邬江兴.基于加权累积和检验的加密流量盲识别算法.软件学报,2013,24(6):1334−1345. http://www.jos.org.cn/1000-9825/4279.htm

英文引用格式: Zhao B, Guo H, Liu QR, Wu JX. Protocol independent identification of encrypted traffic based on weighted cumulative sum test. Ruan Jian Xue Bao/Journal of Software, 2013,24(6):1334−1345 (in Chinese). http://www.jos.org.cn/1000-9825/4279.htm

Protocol Independent Identification of Encrypted Traffic Based on Weighted Cumulative Sum Test

ZHAO Bo, GUO Hong, LIU Qin-Rang, WU Jiang-Xing

(National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China)

Corresponding author: ZHAO Bo, E-mail: chaoxbang@gmail.com

Abstract: A protocol independent identification algorithm is proposed to identify encrypted traffic from both public and private encryption protocols. The randomness of the packet is evaluated by a cumulative test. In addition, results are weighted conflated. A test is performed when every new packet arrived rather than after all packets have received, so that time consumed computation is avoided. The quantity of packets may vary dynamically according to delay and accuracy requirement. Experiments results show that the algorithm achieves accuracy above 90% for SSL and private encryption protocol traffic.

Key words: traffic classification; encrypted traffic identification; cumulative sum test

网络流量识别对于服务质量保证、网络规划建设、网络异常检测等均有重要意义,是进行流量工程、实施 QoS 保障的基础;此外,网络负载建模、流量整形等问题的解决也依赖于有效的流量识别[1].随着互联网的飞速发展,加密协议应用越来越广泛和深入,加密流量的识别作为流量识别的子问题,受到越来越多的重视.

流量识别针对特定的目标,根据被动观察的特征,将网络数据流划分为不同的集合或者将数据流中的某些集合区分出来.流量识别的方法主要包括深入报文检测和深入流检测两大类.由于加密流量的特殊性,其内容不具备固定的特征字段,经信道加密后不改变流量特征,使得传统的业务识别方法难以直接应用于加密流量的普适识别中.目前,已有的识别加密流量的方法主要从加密流量的协议握手特征出发,利用握手及密钥协商阶段明文通信的特性,通过载荷特征匹配或者报文流量指纹匹配进行识别.但这些方法具有共性的缺点,仅能够对特定

* 基金项目: 国家高技术研究发展计划(863)(2009AA01A346); 国家发改委专项(CNGI-09-02-03)
收稿时间: 2011-07-11; 定稿时间: 2012-05-31

加密协议的特定版本有效,无法实现普适识别.此外,一些算法耗时较长,难以实现在线识别.

为实现加密流量的在线普适识别,需分析加密流量区别于非加密流量的特征.从密码学角度来看,加密流量数据为达到安全性需求,会消除所有统计特征,即近似为随机数据.已有研究利用加密数据的高随机性特征识别加密过的恶意软件,或者定位可执行文件数据中加密的部分.但是,这些方法的处理对象和处理环境与本文不同,无法解决加密流量的在线普适识别问题.

鉴于目前对加密流量识别研究的不足,本文提出一种基于加权累积和检验的时延自适应加密流量盲识别算法——EIWCT(encrypted traffic identification based on weighted cumulative sum test),该方法能够实现对包括SSL,SSH 等常用加密协议流量及其他私有加密协议流量的在线识别.其主要思想是:对数据流中的数据报文逐一进行累积和检验,衡量其随机性,根据报文长度对结果进行加权综合.当新报文到达时,无需对所有数据重复耗时较长的累积和检验,而仅需对新到达报文进行累积和检验,并对检验结果进行综合即可.具有快速、普适的特点,适用于骨干链路上的加密流量的在线甄选识别,从而为加密流量的管控,以及用户隐私保护、非法数据监控、网络攻击检测提供基础.

1 相关研究工作

1.1 研究现状

流量识别的传统方法包括基于深入报文检测的识别方法[2]和基于深入流检测的识别方法[3].在加密流量识别上,传统的识别方法面临一系列问题,主要是:

- 基于深入报文检测的方法依赖于对报文头部或载荷内容的匹配分析.加密流量的报文载荷为密文数据,不具备可供匹配的内容特征字段,传统的深入报文检测算法难以直接应用.尽管部分常用加密协议具有固定的通信端口,可通过特定的端口信息识别,但随着随机端口和私有协议的广泛应用,这种检测方法的准确率严重下降[4,5];
- 基于深入流检测的识别方法依赖于单条数据流自身或从属于同一业务的多个数据流间的流量统计特性.对于同样的业务数据,分别通过加密和非加密流量承载,例如 P2P 流量和加密 P2P 流量,其流量特征未有明显区别[6].这使得基于深入流检测的识别方法加密流量也不可行.

目前,已有的加密流量识别方法主要针对协议的连接建立阶段,利用特定的握手及密钥协商明文信息,通过内容特征匹配或者流量特征匹配的方法进行识别.

Bernaille 等人[6]提出了一种识别 SSL 协议流量的方法,其充分利用协议建立连接阶段的信令内容特征来实现 SSL 协议流量的识别.例如,一个连接是否使用了 SSLv2 加密,可以通过连接的第 2 个报文(服务器向户端发送的第 1 个报文)来判断.前面 2bits 内容总为 1 和 0,接下来的 14bit 则指示了 SSL 的长度,第 3 个字节指示消息类型(1 为客户端的 hello 消息,4 为服务器端的 hello 消息).该方法利用 SSL 协议连接建立阶段的明文特征完成识别,并不适用于其他的加密协议.

Alshammari 等人[4,7]对 SSH 流量的识别问题展开研究,提出了一种 SSH 流量识别方法.该方法基于 SSH 协议通过明文信令建立连接的事实,选择了 13 个特征字段和 14 个流属性,其中不包括载荷、IP 地址和端口信息,通过基于包括 C4.5,Naive Bayesian 和 SVM 在内的多种有监督机器学习算法实现对 SSH 协议流量的识别.尽管作者在多种网络上进行了验证,保证了算法的普遍有效性,但是识别的对象并不普遍,局限于 SSH 协议流量.此外,识别过程需要遍历整个数据流,不能够在线识别.

Haffner 等人[8]比较了 AdaBoost、纯真贝叶斯等多种业务识别模型,结果显示,AdaBoost 的效果最好,对 SSH 能达到 86%的识别率和 0.0%假阳性率.不过,该方法需要考察报文载荷的前 64 字节.由于 SSH 的报文加密在握手报文之后,所以对 SSH 流量的识别效果较好.然而,如果加密算法改变,那么识别效果将受到影响.此外,该方法对其他加密应用如 Skype 或者 VPN 隧道并不通用.Baset 等人[9]使用类似的方法,通过握手报文来识别加密的 Skype 流量.该类方法对报文顺序十分敏感,报文乱序对准确率影响较大.此外,识别器需要一定的训练过程,识别耗时较长.

上述方法具有共性的缺点,仅能够对特定加密协议的特定版本有效,识别对象的范围较窄,对未知加密协议不具备识别能力.此外,一些基于机器学习的算法需要训练过程,耗时较长,难以实现在线识别.

1.2 加密流量的特征

加密流量的本质是加密协议交互产生的加密数据,而加密协议的核心是分组加密算法.下面分别介绍分组加密算法的基础知识和加密数据的特征.

加密就是把明文转换为不可辨识的密文的过程,使非授权人无法识别和篡改.为了达到在密文中隐藏明文和密钥信息的目的,在设计加密算法时,会尽力消除密文中所包含的所有特征信息,即让密文序列具备零内容特征.加密流量的零内容特征是由其密文属性本质决定的,其根本特性是信息的隐蔽性,即根据密文难以获取明文及密钥信息,能够抵御各种暴力破解和密码分析攻击.从另一个角度来看,加密流量具备的零内容特征恰恰是其特征,据此设计零内容特征的识别方法,可实现加密流量的判定.

网络上运行的加密协议的核心是分组加密算法,其本质是用有限长的密钥将近似无限长的明文序列加密为密文序列.分组加密算法的一般设计准则是 Shannon 提出的扩散准则和混乱准则[10],基于这两个设计准则,分组加密算法应具有扩散特性和混乱特性.扩散特性指加密算法对密钥或明文的变化是敏感的,包括明文扩散特性和密钥扩散特性.根据严格雪崩准则,明文或密钥任意比特的变化,应导致大约一半密文比特的变化.混淆特性指密钥和明文以及密文之间的依赖关系相当复杂,以至于这种依赖性对密码分析者来说是无法利用的,包括明密文独立性和密文输出随机性.明密文独立性指明文和密文应该是统计独立的,密文输出随机性指对于任意的明文输入,输出的密文序列应该是随机的[11].

在分组密码算法的设计中,通过随机性检测是一个必要和基本的条件,即密码算法输出的密文序列应为近似随机的序列,在严格测试的条件下,不应检测出明显的非随机特征.这是确保密码算法安全性的基础.数据的随机性成为判断其是否为加密数据的依据.Lyda 等人[12]提出了一种基于信息熵值测量的恶意软件识别方法,能够识别出经过加密和压缩处理的恶意软件.作者应用熵测量软件 Bintropy 测量文件数据的信息熵值,通过测量到的信息熵值判断其是否为加密和压缩数据.Hayden[13]研究了寻找可执行文件中加密代码的方法.作者对可执行文件实施随机性检测,根据检测结果定位文件中加密的部分.作者研究了 4 种加密算法 AES,DES,RSA 和 TEA,取得了较好的结果.这些研究均根据加密后的数据具有较高随机性的原理,通过统计性测试来定位文件中加密的部分.

这些方法并不适用于本文的研究环境,它们的处理对象是存储在终端上的数据,即将连续的二进制序列作为一个整体处理.这些方法对数据长度有下限要求,例如文献[14,15]的方法要求密文数据部分长度至少为 100 个双字.此外,上述方法在处理时间上没有限制.而本文研究的对象是网络上传输的流量,以数据报文的形式依次到达,报文长度从几十到上千字节不等.为实现在线识别,处理时间上有严格要求.这些需求,目前的研究成果尚不能满足.

2 EIWCT 算法

前文分析了当前与加密流量识别有关的研究存在的问题以及加密流量的特征.不难看出,加密流量数据具有随机性高的特点,并且以离散不等长报文的形式到达.基于在线普适的识别需求,本文提出了一种基于加权累积和检验的时延自适应加密流量盲识别算法 EIWCT.其主要思想是:对数据流中的报文数据序列逐一进行累积和检验,衡量其随机性,根据报文长度对结果进行加权综合.当新报文到达时,无需对所有数据重复耗时较长的累积和检验,而仅需对新到达报文进行累积和检验,并对检验结果进行综合即可.

2.1 累积和检验

累积和检验是一种假设检验,即假设检验对象是随机二进制序列,计算以多大的概率接受该假设.文献[16]给出了随机二进制序列(以下简称随机序列)的定义,由独立的伯努利随机变量构成的二进制序列,其中,$P\{\varepsilon_i=0\}=P\{\varepsilon_i=1\}=1/2$.随机序列具有以下特性:

1) 对称性(uniformity):0 和 1 的数量应近似相等,数学期望相同,均为 $n/2$.n 为序列长度;

2) 扩展性(scalability):一个随机序列的子序列也应该是随机的.

对随机序列来说,构成序列的 0 和 1 的数量在概率上相等.如果把随机序列看做一段连续的指令,每一位代表一个动作,0 代表向左走,1 代表向右走,起始点为原点,那么当整个序列的指令执行完毕时,最终的位置不应该离原点太远,整个过程中偏离原点的最远距离也不应该很大.如果偏离原点的最远距离过大,说明序列的一部分 0 和 1 的差异过大,序列不随机.这就是累计和检验的基本原理.下面在基本原理的基础上,以随机序列为对象,从理论上分析偏离原点的最远距离与序列长度的关系,并据此设计检验算法.

2.1.1 理论分析

首先定义如下符号:

- ε_i:序列的第 i 比特;
- η_i:序列第比特的变形,$\eta_i=2\varepsilon_i-1$;
- η:待检序列,也就是初始序列变形后的序列 $\eta=\eta_1\eta_2\ldots\eta_n$;
- n:序列的比特长度;
- S_k:序列前 k 位之和,$S_k=\sum_{i=1}^{k}\eta_i$;
- z_n:S_k 的绝对值最大值,$z_n=\max_{1\leqslant k\leqslant n}|S_k|$.

定理 1.

$$P\{S_n=k\}=\begin{cases}\binom{n}{(n-k)/2}2^{-n}, & \text{当}k\equiv n\ (\mathrm{mod}\,2)\text{时}\\ 0, & \text{其他}\end{cases},$$

其中,$k=-n,-n+1,\ldots,n;n=1,2,\ldots$

证明:根据排列组合,显然有:

开始

$$P\{S_{2n}=2k\}=\binom{2n}{n-k}2^{-2n} \tag{1}$$

其中,$k=-n,-n+1,\ldots,n;n=1,2,\ldots$

$$P\{S_{2n}=2k+1\}=\binom{2n+1}{n-k}2^{-2n-1} \tag{2}$$

其中,$k=-n-1,-n,\ldots,n;n=1,2,\ldots$

由公式(1)、公式(2)得:

$$P\{S_n=k\}=\begin{cases}\binom{n}{(n-k)/2}2^{-n}, & \text{当}k\equiv n\ (\mathrm{mod}\,2)\text{时}\\ 0, & \text{其他}\end{cases},$$

其中,$k=-n,-n+1,\ldots,n;n=1,2,\ldots$ □

定理 2. 对任意整数 $b\geqslant 0,-b\leqslant v\leqslant b$,有

$$p_n(b,v)=P\{z_n<b,S_n=v\}=\sum_{k=-\infty}^{\infty}q_n(4kb+v)-\sum_{k=-\infty}^{\infty}q_n((4k+2)b-v),$$

其中,$q_n(j)=P\{S_n=j\}$.

证明:根据定理 1,有:

- $q_0(j)=P\{S_0=j\}=\begin{cases}1, & j=0\\ 0, & j\neq 0\end{cases}$;
- $q_n(j)=q_n(-j)$;

- $q_n(j)=\frac{1}{2}q_{n-1}(j-1)+\frac{1}{2}q_{n-1}(j+1)$.

(1) 当 n=0 时,根据其物理含义,显然有:

- 左侧= $p_0(b,v)=\begin{cases}1, & \text{当}v=0\text{且}b>0\text{时}\\ 0, & \text{其他}\end{cases}$;
- 对于等号右侧:

  当 $v\neq0$ 时,因为$|v|<b$,所以 $4kb+v\neq0$ 且$(4k+2)b-v\neq0$,所以等号右侧=0−0=0;

  当 v=0 时:

  ① b=0,$4kb+v$=0,$(4k+2)b-v$=0,等号右侧=1−1=0;

  ② b>0,

$$\begin{aligned}\text{等号右侧}&=\sum_{k=-\infty}^{\infty}q_0(4kb)-\sum_{k=-\infty}^{\infty}q_0(4k+2)b\\&=q_0(0)-\sum_{k=-\infty}^{\infty}q_0(4k+2)b\\&=1-0\\&=1.\end{aligned}$$

综上,等号左侧=等号右侧.

(2) 当 $n\neq0$ 时,易证 $p_n(0,v)$=0=右侧.下面证明 b>0 的情况.

假设当 $n=n_0$ 时,等式成立.

$$P_{n_0}(b,v)=\sum_{k=-\infty}^{\infty}q_{n_0}(4kb+v)-\sum_{k=-\infty}^{\infty}q_{n_0}((4k+2)b-v),$$

则当 $n=n_0+1$ 时,

$$\begin{aligned}\text{右侧}&=\sum_{k=-\infty}^{\infty}q_{n_0+1}(4kb+v)-\sum_{k=-\infty}^{\infty}q_{n_0+1}((4k+2)b-v)\\&=\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}(4kb+v-1)+\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}(4kb+v+1)-\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}((4k+2)b-v-1)-\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}((4k+2)b-v+1)\\&=\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}(4kb+v-1)-\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}((4k+2)b-v+1)+\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}(4kb+v+1)-\frac{1}{2}\sum_{k=-\infty}^{\infty}q_{n_0}((4k+2)b-v-1)\\&=\frac{1}{2}p_{n_0}(b-1,v-1)+\frac{1}{2}p_{n_0}(b+1,v+1)\\&=p_{n_0+1}(b,v)\\&=\text{左侧}.\end{aligned}$$

综上所述,对所有 n,等式均成立. □

定理 3. 对任意整数 $b\geqslant0$,有

$$P\{z_n<b\}=\sum_{k=-\infty}^{\infty}P\{(4k-1)b<S_n<(4k+1)b\}-\sum_{k=-\infty}^{\infty}P\{(4k+1)b<S_n<(4k+3)b\}.$$

证明:对满足$-b\leqslant u\leqslant v\leqslant b$ 的整数 u,v,由定理 1,有

$$P\{M_n<b,u<S_n<v\}=\sum_{k=-\infty}^{\infty}P\{u+4kb<S_n<v+4kb\}-\sum_{k=-\infty}^{\infty}P\{2b-v+4kb<S_n<2b-u+4kb\}.$$

显然有 $u<b,v<b$.令 $u=v=b$,有

$$\begin{aligned}P\{z_n<b,-b<S_n<b\}&=\sum_{k=-\infty}^{\infty}P\{-b+4kb<S_n<b+4kb\}-\sum_{k=-\infty}^{\infty}P\{2b-b+4kb<S_n<2b+b+4kb\}\\&=\sum_{k=-\infty}^{\infty}P\{(4k-1)b<S_n<(4k+1)b\}-\sum_{k=-\infty}^{\infty}P\{(4k+1)b<S_n<(4k+3)b\}.\end{aligned}$$

□

2.1.2 检验步骤

由 z_n 的物理含义可知:

$$P(z_n < z) = P\left(\frac{\max|S_n|}{\sqrt{n}} < z\right) \tag{3}$$

根据中心极限定理,有

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sqrt{n}} \leqslant z\right) = \Phi(z) \tag{4}$$

其中,$\Phi(z)$为标准正态函数:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \mathrm{e}^{-u^2/2}\mathrm{d}u.$$

综合公式(3)、公式(4)及定理 3,累积和检验的具体方法为:

1) 将原始二进制序列 $\varepsilon=\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$ 中的所有 0 替换为 −1,1 不变.即生成新的序列 $\eta=\eta_1\eta_2\ldots\eta_n$,其中,$\eta_i=2\varepsilon_i-1$;

2) 计算新序列的前 k 项和

$$S_k = \sum_{i=1}^{k}\eta_i.$$

3) 计算最大偏移值 z

$$z=\max_{1\leqslant k\leqslant n}|S_k|.$$

4) 计算检验值 $I(z)$

$$I(z) = 1 - \sum_{k=\left(\frac{-n}{z}+1\right)\Big/4}^{\left(\frac{n}{z}-1\right)\Big/4}\left[\Phi\left(\frac{(4k+1)z}{\sqrt{n}}\right) - \Phi\left(\frac{(4k-1)z}{\sqrt{n}}\right)\right] + \sum_{k=\left(\frac{-n}{z}-3\right)\Big/4}^{\left(\frac{n}{z}-1\right)\Big/4}\left[\Phi\left(\frac{(4k+3)z}{\sqrt{n}}\right) - \Phi\left(\frac{(4k+1)z}{\sqrt{n}}\right)\right],$$

其中,$\Phi(z)$为标准正态分布函数:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \mathrm{e}^{-u^2/2}\mathrm{d}u.$$

2.2 序列检验值的分布

本节讨论不同类型序列的累积和检验值分布.对于一个序列,判断其是否为随机序列的依据是其检验值的分布情况,这是本文识别算法设计的出发点和基础.

序列的累积和检验值 $I(z)$的分布记为 X,其概率密度函数记为 $f(x)$.当序列完全随机时,X 应为[0,1]内的均匀分布;当序列不随机时,受其非随机性特征影响,X 具有不规则的特定分布,情况较为复杂.

根据经验值推断,对非随机序列的检验,检验值 $I(z)$会整体偏小.

下面按照检验值 $I(z)$的获取过程证明这个假设:首先证明,当二进制序列的码元等概率时,检验值 $I(z)$计算过程第 3 步的最大偏移值最小;接着证明,当最大偏移值最小时,检验值 $I(z)$最大;最后证明,当二进制序列的码元不等概,即为非随机序列时,其检验值 $I(z)$小于最大值.

引理 1. 对于二进制序列来说,当 $P\{\varepsilon_i=0\}=P\{\varepsilon_i=1\}=1/2$ 时,最大偏移值 z_r 最小.

证明:考察序列 ε,$P\{\varepsilon_i=0\}=p$,$P\{\varepsilon_i=1\}=q$.其前 k 项和与最大偏移值分别定义为 $S_k(p,q)$和 $z(p,q)$.

由于二元码元的对称性,显然有

$$P\{S_k(p,q)=x\}=P\{S_k(q,p)=-x\}.$$

因此,

$z(p,q)$在 $p=q=1/2$ 时取得极值.

根据大偏差定理,有

$$\lim_{n\to\infty} n^{-1}\log P\{S_n \geqslant n\alpha\} = -\frac{\alpha}{2}\log\frac{q(1+\alpha)}{p(1-\alpha)} + \frac{1}{2}\log\frac{4pq}{1-\alpha^2},$$

其中,$0<\alpha<1$.

令 $M(q) = -\frac{\alpha}{2}\log\frac{q(1+\alpha)}{p(1-\alpha)} + \frac{1}{2}\log\frac{4pq}{1-\alpha^2}$,

对 $M(q)$求导,有

$$M'(q) = \frac{1-\alpha}{2q}.$$

由于 $0<\alpha<1, 0<q<1$,因此 $M'(q)>0$,即 $M(q)$为单调递增函数.

因此,最大偏移值 $z(p,q)$在 $p=q=1/2$ 取得最小值.

对于随机序列,码元等概出现,即 $P\{\varepsilon_i=0\}=P\{\varepsilon_i=1\}=1/2$,因此随机序列的最大偏移值 z_r 最小,即

$$z_r=\min\{z\}.$$ □

引理 2. $I(z_1)>I(z_2)$,其中,$0<z_1<z_2<n$.

证明:

$$\begin{aligned}
I'(z) = &\sum_{k=\left(\frac{-n}{z}-3\right)/4}^{\left(\frac{n}{z}-1\right)/4}\left[(4k'(z)z+4k+3)\exp\left(-\frac{(4k+3)^2z^2}{2n}\right)-(4k'(z)z+4k+1)\exp\left(-\frac{(4k+1)^2z^2}{2n}\right)\right]-\\
&\sum_{k=\left(\frac{-n}{z}+1\right)/4}^{\left(\frac{n}{z}-1\right)/4}\left[(4k'(z)z+4k+1)\exp\left(-\frac{(4k+1)^2z^2}{n}\right)-(4k'(z)z+4k-1)\exp\left(-\frac{(4k-1)^2z^2}{n}\right)\right]\\
= &\sum_{k=\left(\frac{-n}{z}-3\right)/4}^{\left(-\frac{n}{z}+1\right)/4}\left[(4k'(z)z+4k+3)\exp\left(-\frac{(4k+3)^2z^2}{2n}\right)-(4k'(z)z+4k+1)\exp\left(-\frac{(4k+1)^2z^2}{2n}\right)\right]\\
> &\sum_{k=\left(\frac{-n}{z}-3\right)/4}^{\left(-\frac{n}{z}+1\right)/4}\left\{(4k'(z)z+4k+1)\left[\exp\left(-\frac{(4k+3)^2z^2}{2n}\right)-\exp\left(-\frac{(4k+1)^2z^2}{2n}\right)\right]\right\}.
\end{aligned}$$

由于 k 的取值范围为 $\left[\frac{1}{4}\left(-\frac{n}{z}-3\right),\frac{1}{4}\left(-\frac{n}{z}+1\right)\right]$,且 $z<<n$,所以$|4k+3|<|4k+1|$,因此有

$$\exp\left(-\frac{(4k+3)^2z^2}{2n}\right)-\exp\left(-\frac{(4k+1)^2z^2}{2n}\right)>0.$$

又因在区间 $\left[\frac{1}{4}\left(-\frac{n}{z}-3\right),\frac{1}{4}\left(-\frac{n}{z}+1\right)\right]$ 内 $k'(z)>0$,

因此有 $\sum_{k=\left(\frac{-n}{z}-3\right)/4}^{\left(-\frac{n}{z}+1\right)/4}\left\{(4k'(z)z+4k+1)\left[\exp\left(-\frac{(4k+3)^2z^2}{2n}\right)-\exp\left(-\frac{(4k+1)^2z^2}{2n}\right)\right]\right\}>0,$

继而有 $I'(z)\geqslant 0$. □

定理 4. 序列检测值的分布为 X,其概率密度函数为 $f(x)$.当序列完全随机时,有

$$f(x)=\begin{cases}0, & x<0\\ 1, & 0\leqslant x\leqslant 1.\\ 0, & x>1\end{cases}$$

当序列不随机时,其概率密度函数为

$$f(x)=\begin{cases}0, & x<0\\ g(x), & 0\leqslant x\leqslant 1.\\ 0, & x>1\end{cases}$$

$\exists\delta>0$,使得 $\int_{1-\delta}^{1}1-g(x)>0.$

证明:对随机序列的最大偏移值 z_r 和非随机序列的最大偏移值 z_n,根据引理 1 和引理 2,有:

- z_n 依概率大于 z_r;
- $I(z_n)$依概率小于 $I(z_r)$.

由于 $I(z_r)$和 $I(z_n)$的值域为[0,1],且 $I(z_r)$的分布为均匀分布,因此,$\exists\delta>0$,使得 $P\{I(z_n)<1-\delta\}>P\{I(z_r)<1-\delta\}$,继而有

$$\int_{0}^{1-\delta}1-g(x)>0.$$

所以,

$$\int_{1-\delta}^{1}f(x)-g(x)>0.$$ □

由定理 4 可知,尽管加密和非加密序列的累积和检验值 $I(z)$范围均为[0,1],但是非加密序列检验值的分布更靠近 0,而加密序列的检验值则是均匀分布.据此,我们设计了识别算法.

2.3 算法描述

根据前面的分析,本节提出一种基于加权累积和检验的时延自适应加密流量盲识别算法,用于网络加密流量的在线检测.其原理是:对数据流的报文逐一进行累积和检验,根据时延要求确定处理的报文数量;同时,根据报文长度实施样本数量加权.检验步骤分两步:第 1 步对报文进行加权累积和检验,生成检验值;第 2 步根据检验值分布输出识别结果.

本质上,累积和检验属于假设检验的一种,即提出一个零假设,根据计算结果以一定概率接受该假设.首先构造零假设和备择假设:

零假设:检验值 $I(z)$符合分布函数为 $F(z)$的分布.

备择假设:检验值 $I(z)$不符合分布函数为 $F(z)$的分布,其中, $F(z)=\int_{-\infty}^{z}f(x)\mathrm{d}x.$

检验值 $I(z)$是[0,1]之间的实数,在确定了置信度水平α后,可判断单个样本是否通过检验.对多个样本的检验将获得多个检验值.如果零假设成立的话,通过检验的样本应达到一定的比例 η_0,其计算方法公式如下:

$$\eta_0=p-3\sqrt{\frac{p(1-p)}{m}},$$

其中,$p=1-\alpha$,m 为样本数量.

根据置信度水平α计算拒绝阈值η_0.需要注意的是,在实际流量的检验中,加密流量中可能包含部分非加密数据,例如特定的头部等,因此,需要根据实际需要适当调整η_0 的取值.如果通过检验的样本比例大于阈值η_0,则接受零假设,认为该数据流的内容是加密的;反之,则接受备择假设.

算法 1. 加权累积和检验算法.

1) 根据报文协议,将报文 p 去除报头获得载荷内容;
2) 将载荷拆分为若干待测序列;
3) 记录样本序列数量 $n=l/L_s$.其中,l 为待测序列长度,L_s 为样本长度;
4) 对每一样本序列进行累积和检验,并记录检验值 $I(z)$;
5) 比较检验值与置信度水平,如果 $I(z)>\alpha$,则 b++.其中,b 为通过检测的样本数量.

本文采用的是假设检验的思想,本质上是一种概率检测.由于偶然事件的影响,仅通过一次检验就下结论是存在偏差的,所以通过增加样本数量来提高检测准确率.即,在数据流中对多个报文实施累积和检验,并将检测结果实时综合处理.

算法 2. 检验值分布检测算法.

1) For i=1 to N do

① 接收从属于同一数据流 s 的第 i 个报文 p_i;

② 调用算法 1,并记录 n_i,b_i.n_i 和 b_i 分别为第 i 个报文 p_i 中的样本数量和通过检测的样本数量,下同;

2) 计算 $n_t=\sum_{i=1}^{N}n_i$;

3) 计算 $\eta_t=\frac{1}{n_t}\sum_{i=1}^{N}b_i$;

4) 计算 $\eta_0=p-3\sqrt{\frac{p(1-p)}{n_t}}-\Delta_\eta$;

5) 如果 $\eta_t>\eta_0$,则返回 true;否则,返回 false.

N 的取值视实际需求而定.随着 N 的增加,显然有准确率的提高;同时,由于需接收更多的报文,将导致时延增加以及存储开销的增加.早期的检测有利于更好地控制和解决问题,也有利于网络管理员能够更有效地实施管理操作.

2.4 算法复杂度分析

EIWCT 算法由加权累积和检验与检验值分布检测两部分组成,它们的计算复杂度为 $O(n)$和 $O(1)$,所以 EIWCT 算法的计算复杂度为 $O(n)$.

处理一个数据流需要存储单个数据,需要的存储开销为 $O(l)$.处理整条链路的数据,存储开销为 $O(l)\cdot O(M)$,M 为链路中并发流数量.

3 EIWCT 算法实验仿真及评估

本节分别以加密和非加密流量为对象,利用第 2 节给出的算法,研究算法的区分度、准确率等方面的性能.

3.1 实验环境

作为研究对象的加密流量取自网上银行网站数据(SSLv3 协议加密)和私有加密协议(AES 加密算法)数据,非加密流量取自普通新闻类网站数据、在线视频类网站数据、在线音频类网站数据和邮箱网站数据.对每一类数据,分别取 40K 字节,并按测试需求分割为不同长度的样本,以考察不同情况下的性能.样本数量取 200,显著性水平α=0.1.

3.2 样本长度对检验值的影响

考察不同样本长度条件下,加密数据与非加密数据检验结果的差异.样本长度分别取 100,200,400,800 和 1 600.累积和检验对样本序列长度的要求为至少 100 比特.长度过短,在统计检验中将产生较大的误差;长度过长,则检验所需数据过多,导致时延和存储开销增加.分别对各类样本实施检验,进行多次测量,结果取平均值.实验结果如图 1 所示.

从图 1 可以看出,在样本长度为 100 和 200 时,加密(网上银行和私有协议)和非加密(新闻网站、在线视频、电子邮件和在线音频)流量的检验值相差不多.当样本长度超过 400 后,加密和非加密流量的检验值开始呈现较大的差异.样本长度越长,这个差异越明显.分析原因,在二元码元序列中,长度在 200 以下的序列样本,由于长度较短,还不能够很好地体现统计特性,难以对随机序列和非随机序列做出明显的区分.随着样本长度的增加,样本所具备的特性越来越明显,这也导致非加密流量的检验值在样本长度超过 400 后急剧减少.但是在实际处理中,受处理时延和存储开销的限制,样本长度不能无限放大.以下的讨论中,取 400 作为典型的样本长度.

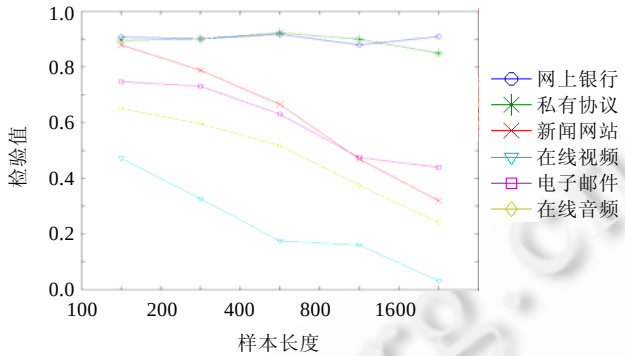

Fig.1 Differentiabilities of test value under various sample length

图 1 不同样本长度下,检验值的区分性能

3.3 样本数量对检验值的影响

当样本长度为典型样本长度 400 时,讨论在不同样本数量情况下,检验值的变化情况.进行一次检验得到的各类流量检验值,并应用算法 2 计算其 η_t 值,作为判断依据.

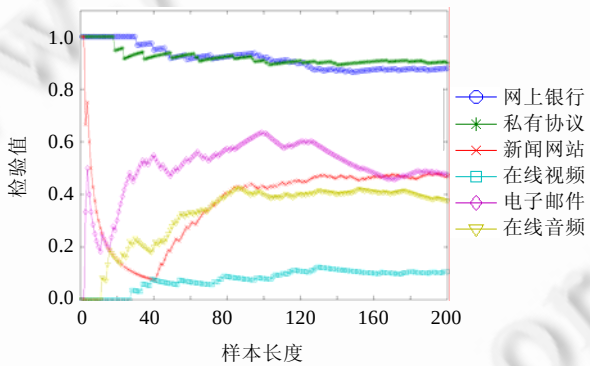

Fig.2 Differentiabilities of test value under various sample size

图 2 不同样本数量下,检验值的区分性能

可以看到,在样本数量为 10 时,各类流量根据算法 2 计算出的 η_t 值已趋于稳定,可以做出有效的判别.由于样本长度为 400,即单个样本大小为 400 比特、50 字节,因此接收到约 0.5K 字节的数据时即可做出有效判断.

3.4 样本数量对准确率的影响

在样本长度为 400 条件下,讨论样本数量与准确率的关系.设定判断阈值 η_0=0.8,超过该阈值则判定为加密流量,否则判定为非加密流量.重复多次测试,准确率如图 3 所示.

从图 3 可以看出:在样本数超过 10 以后,准确率达到 80%以上;样本数超过 13 以后,准确率提高到 90%以上;在样本数量超过 20,即数据量超过 1K 字节后,准确率稳定在 90%以上,并随着样本数量的增加继续增加.在准确率要求不高的场景中,较少的样本需求可减少识别时延;而当识别时延较为宽松时,可接收更多的样本,获得更高的准确率.

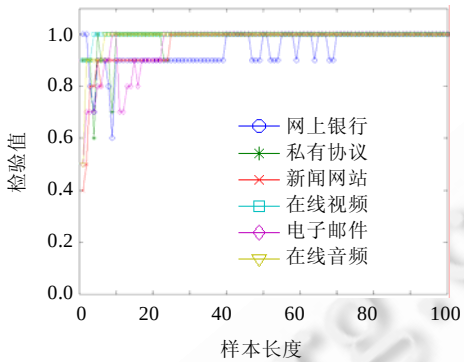

Fig.3 Sample size influence on accuracy rate

图 3 不同样本数量下,准确率的变化

4 结 论

针对加密数据不具备普遍内容特征、难以实现加密流量普适识别的问题,本文提出了一种基于加权累积和检验的时延自适应加密流量盲识别算法EIWCT.对网络数据报文实施累积和检验,根据报文长度加权综合,实现识别结果的在线输出.根据处理时延的要求,可动态调整检验报文的数量,在时延限制内达到准确率最大化.

基于常用的加密协议 SSL 和私有加密协议,对比未加密的网页数据及视频、音频和文本数据等,进行了实验仿真.结果表明,EIWCT 算法可实现在线处理,在序列长度为 400 比特、数据量超过 1K 字节后,识别准确率达到 90%以上.此外,该算法具有较强的灵活性,可实时调整样本长度和用于判断的数据量.随着二者的增加,实现准确率的进一步提高.

References:

[1] Alshammari R, Zincir-Heywood AN. A flow based approach for SSH traffic detection. In: Proc. of the IEEE Int'l Conf. on Systems, Man and Cybernetics (ISIC). 2007. 296−301. [doi: 10.1109/ICSMC.2007.4414006]

[2] Yu Q, Huo HW. Algorithms improving the storage efficiency of deep packet inspection. Ruan Jian Xue Bao/Journal of Software, 2011,22(1):149−163 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3724.htm [doi: 10.3724/SP.J.1001.2011.03724]

[3] Xu P, Lin S. Internet traffic classification using C4.5 decision tree. Ruan Jian Xue Bao/Journal of Software, 2009,20(10): 2692−2704 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3444.htm [doi: 10.3724/SP.J.1001.2009.03444]

[4] Alshammari R, Zincir-Heywood AN. Generalization of signatures for SSH encrypted traffic identification. In: Proc. of the Computational Intelligence in Cyber Security. 2009. 167−174. [doi: 10.1109/CICYBS.2009.4925105]

[5] Bernaille L, Teixeira R, Akodkenou I, Soule A, Salamation K. Traffic classification on the fly. SIGCOMM Computer Communication Review, 2006,36(2):23−26. [doi: 10.1145/1129582.1129589]

[6] Bernaille L, Teixeira R. Early recognition of encrypted applications. In: Proc. of the 8th Int'l Conf. on Passive and Active Network Measurement (PAM 2007). Louvain-la-Neuve, 2007. 165−175. [doi: 10.1007/978-3-540-71617-4_17]

[7] Alshammari R, Zincir-Heywood AN. Investigating two different approaches for encrypted traffic classification. In: Proc. of the 2008 Sixth Annual Conf. on Privacy, Security and Trust. 2008. 156−166. [doi: 10.1109/PST.2008.15]

[8] Haffner P, Sen S, Spatscheck O, Wang DM. ACAS: Automated construction of application signatures. In: Proc. of the ACM SIGCOMM Workshop on Mining Network Data. 2005. 197−202. [doi: 10.1145/1080173.1080183]

[9] Baset SA, Schulzrinne HN. An analysis of the skype peer-to-peer Internet telephony protocol. In: Proc. of the IEEE Infocom 2006. 2006. 1−11. [doi: 10.1109/INFOCOM.2006.312]

[10] Lai XJ, Massey JL, Murphy S. Markov ciphers and differential cryptanalysis. In: Proc. of the Advances in Cryptology (EUROCRYPT'91). Berlin: Springer-Verlag, 1991. 17−38. [doi: 10.1007/3-540-46416-6_2]

[11] Wu WL, Feng DG, Zhang WT. Design and Analysis of Block Cipher. Beijing: Tsinghua University Press, 2000 (in Chinese).

[12] Lyda R, Hamrock J. Using entropy analysis to find encrypted and packed malware. IEEE Security & Privacy, 2007,5(2):40−45. [doi: 10.1109/MSP.2007.48]

[13] Hayden WJ. Locating encrypted data hidden among non-encrypted data using statistical tools [MS. Thesis]. Air Force Air University, 2007.

[14] Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, Levenson M, Vangel M, Banks D, Heckert A, Dray J, Vo S. A statistical test suite for random and pseudorandom number generators for cryptographic applications. NIST, 2010. http://csrc.nist.gov/groups/ST/toolkit/rng/documentation_software.html

[15] NIST. NIST/SEMATECH e-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/

[16] Durrett R. Probability: Theory and Examples. 3rd ed., Belmont: Duxbury, 2004.

附中文参考文献:

[2] 于强,霍红卫.一组提高存储效率的深度包检测算法.软件学报,2011,22(1):149−163. http://www.jos.org.cn/1000-9825/3724.htm [doi: 10.3724/SP.J.1001.2011.03724]

[3] 徐鹏,林森.基于 C4.5 决策树的流量分类方法.软件学报,2009,20(10):2692−2704. http://www.jos.org.cn/1000-9825/3444.htm [doi: 10.3724/SP.J.1001.2009.03444]

[11] 吴文玲,冯登国,张文涛.分组密码设计与分析.北京:清华大学出版社,2000.

赵博(1981－),男,吉林公主岭人,博士生,主要研究领域为流量识别,网络管控.
E-mail: chaoxbang@gmail.com

刘勤让(1975－),男,博士,副教授,主要研究领域为网络业务识别与控制.
E-mail: qinrangliu@gmail.com

郭虹(1975－),女,博士生,讲师,主要研究领域为复杂网络,网络拓扑建模.
E-mail: guo_hong@163.com

邬江兴(1953－),男,教授,博士生导师,中国工程院院士,主要研究领域为信息网络,交换技术.
E-mail: wujiangxing@mail.ndsc.com.cn