基于随机性检测的链路层加密数据盲识别方案

吴 杨[1] 马云飞[2] 王 韬[1] 邢 萌[1]
(军械工程学院信息工程系 石家庄 050003)[1] (郑州大学信息管理系 郑州 450001)[2]

摘 要 为识别链路层加密数据,构建了以块内频数检测为主的链路层加密数据盲识别方案。针对分块长度影响块内频数检测识别率的问题,提出了基于方差原理的比特序列分块长度值选择方案。针对块内频数检测对长度较短的比特序列的识别能力有限的问题,提出了基于随机抽样原理的比特序列信息提取方法,以提高对未加密比特序列的识别率。最后,以某无线网络链路层数据的识别为例,对提出的方案进行了验证。结果表明,提出的方案对链路层加密数据具有较高的识别率,相关成果可为进一步的协议识别技术研究打下基础。

关键词 链路层,加密数据,块内频数检测,方差原理,随机抽样

中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.042

Independent Identification of Encrypted Data in Data Link Layer Based on Randomness Test

WU Yang[1] MA Yun-fei[2] WANG Tao[1] XING Meng[1]
(Department of Information Engineering, Ordnance Engineering College, Shijiazhuang 050003, China)[1]
(Department of Information Management, Zhengzhou University, Zhengzhou 450001, China)[2]

Abstract To identify the encrypted data of data link layer, an independent identification scheme was proposed by mainly using the technique of frequency test within a block. As the identification rate of frequency test within a block is impressionable to the size of the block, a block size chosen scheme was provided based on the principle of variance. Furthermore, as the identification ability of frequency test within a block is also limited to short bit sequence, random sampling was introduced into its information extraction process to heighten the identification rate. Eventually, by taking the link layer bit sequence of a wiriness network as the identification object, the identification rate of the proposed schemes was verified. Experimental results demonstrate that the proposed schemes have high rates to identify encrypted data of data link layer and the correlative research paves the way for the further protocol identification research.

Keywords Data link layer, Encrypted data, Frequency test within a block, Variance principle, Random sampling

1 引言

传统网络协议识别主要针对应用层协议[1],且多为已知网络中的未加密数据。在日益严峻的网络安全形势下,对网络中传输的数据进行加密是确保数据安全传输的有效手段,越来越多的网络应用为确保其信息的机密性,均在链路层对数据进行了加密处理[2-4],而识别链路层加密数据可为进一步开展上层协议识别及网络安全防护技术研究提供支持。当前,在加密流量分析方面,其主要研究内容是对加密流量所属的上层协议类型进行分析。Charles 等[5]提出了通过加密后保持不变的特征(数据包字节数、持续时间和流方向)来研究识别加密流量所属的应用协议,但并未给出判别数据是否加密的方法且方案的针对性不强。文献[6]则提出了一种混合多级的加密流量分类识别方法,其方案以签名匹配的方法对SSL/TLS的流量进行了分类识别,而后采用统计分析方法确定加密流量所属的具体应用协议,但仍无法有效用于对未知协议的识别。针对加密流量识别问题,文献[7]利用加密数据的随机性特点,提出了基于加权累积和检验的自适应加密流量盲识别算法,其特点是对网络报文逐一实施累积和检验,并将结果进行加权综合,以提高对加密流量的识别率。文献[7]提出的算法在执行过程中,需首先对网络报文进行归类预处理,将属性相同的报文归类到相同的聚类当中,其识别率依赖于参与累积和检验的报文需具有相似的属性,该条件使其无法有效应用于对单条报文数据的识别,同时也降低了其算法的适用性。综上所述,针对链路层加密数据识别的相关公开研究成果较少,可用的方法主要是基于特征串匹配等算法,查找数据中是否存在特定的模式串,该类方法只适用于对已知网络中的链路层加密数据进行识别,且需获得协议特征,如何有效实现对已知及未知网络的链路层加密数据进行识别,还需进一步深入研究。

未加密链路层数据的特征字段及用户数据使得其随机统计特性与加密链路层数据的随机统计特性存在一定的差异性。本文以此为突破口,将随机性检测方法用于对链路层数据比特序列的随机性度量,将其比特序列的随机性差异信息

到稿日期:2013-10-02 返修日期:2013-12-19 本文受军内科研项目(YJJXM12033)资助。

吴 杨(1985-),男,博士生,主要研究方向为网络协议识别,E-mail:baiyanwy@163.com;王 韬(1964-),男,教授,博士生导师,主要研究方向为网络安全、密码学。

作为识别其是否加密的主要依据。其次，针对块内频数检测的参数选择问题，将方差原理引入到分块长度的选择过程中。同时，针对块内频数检测方法对比特序列长度敏感的特性，将随机抽样原理引入到比特序列的信息提取过程中，以提高对未加密比特序列的识别率。

本文第2节论述了随机性检测的相关理论与方法，分析了码元频数、游程及块内频数检测用于链路层加密数据的比特序列识别存在的问题；第3节阐述了优化块内频数检测参数的思路，并给出了对应的方案；第4节对实验结果进行了分析；最后对相关工作进行了总结。

2 相关理论与方法

随机性检测建立在假设检验的基础上，其基本思想为小概率原理，即如果认为某个假设是真实的，则在这个假设下发生的概率很小的事件在一次试验中是不可能发生的，如果这一事件发生，则假设的真实性将受到质疑。在该思想基础上，首先给出两个假设：原假设 H_0（序列随机）、备择假设 H_1（序列非随机）。然后，构造在原假设为真时分布已知的统计量。其次，根据已知的分布，给定显著性水平 a 的值，参照备择假设，令 $P(\text{拒绝 } H_0 \mid H_0 \text{ 真})=a$，确定统计量的拒绝域。

保持信息内容不被非授权者获取及机密性的实现方式可从物理保护到采用让数据无法理解的数学算法[8]。同未加密数据相比，已加密数据在统计学上体现出了更大的随机特性。在度量数据的随机性方面，已有众多的随机性检测项目和方法，类似的随机性检测方法更是多达200多种[9]，广泛用于检测密码算法及序列的统计特性[10,11]。其中，应用最为广泛的是NIST于2001年5月公布的FIPS140-2标准中定义的用于密码系统安全性度量的诸多随机性检测方案[12]，及NIST于2010年4月又公布的应用于测试随机数及伪随机数生成器性能的随机性检测标准SP800-22 revla（Special Publication 800-22 Revision 1a）[13]。

在随机性检测研究领域，相关研究热点主要集中于如何提高众多检测方法的运行效率及度量检测方法间的相关性。SP800-22 revla标准中的码元频数、游程和块内频数检测的基本原理及实现过程如下。

2.1 码元频数检测

码元频数检测又称单比特检测，主要用于实现对整个比特序列中比特0和1出现的频数进行检测，以评估序列是否满足随机性要求。为便于描述，定义长度为 n 的比特序列为 $E=\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$。检测过程首先基于 $x_i=2\varepsilon_i-1$ 变换，构造新的序列 $X=x_1,x_2,\cdots,x_n$，并统计序列 X 的二项式和 $S_n=x_1+x_2+\cdots+x_n$，规格化 S_n 得 $\frac{S_n}{\sqrt{n}}$，而 $\frac{S_n}{\sqrt{n}}$ 接近于标准的正态分布函数，即满足 $\lim\limits_{n\to\infty}P(\frac{S_n}{\sqrt{n}}\leqslant z)=\phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-u^2/2}du$。对于统计量 $S_{obs}=\frac{|S_n|}{\sqrt{n}}$ 而言，有如下等式成立：

$$\begin{aligned}&\lim_{n\to\infty}P(S_{obs}=\frac{|S_n|}{\sqrt{n}}\leqslant z)\\&=1-\frac{1}{\sqrt{2\pi}}(\int_{-\infty}^{-z}e^{-u^2/2}du+\int_{z}^{+\infty}e^{-u^2/2}du)\\&=2\phi(z)-1\end{aligned}\tag{1}$$

SP800-22 revla将码元频数检测的统计门限值 P-value 定义为：

$$\begin{aligned}P\text{-value}&=2[1-\phi(S_{obs})]=2(1-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{S_{obs}}e^{-u^2/2}du)\\&=\frac{2}{\sqrt{2\pi}}\int_{S_{obs}}^{+\infty}e^{-u^2/2}du=\frac{2}{\sqrt{\pi}}\int_{\frac{S_{obs}}{\sqrt{2}}}^{+\infty}e^{-t^2}dt\\&=erfc(\frac{S_{obs}}{\sqrt{2}})\end{aligned}\tag{2}$$

当门限值 P-value$\geqslant$0.01时即可认为序列是随机的。

2.2 游程检测

游程为比特序列的子序列，由连续的0或1分别构成0游程或1游程。游程检测在码元频数检测后进行，用于评估序列中的游程数是否满足随机性要求，主要对比特序列中的连续特征位进行检测。游程检测定义 $V_n(obs)$ 为游程总数即0游程和1游程之和，$\pi=\frac{\sum_{j=1}^{n}\varepsilon_j}{n}$ 为序列 ε 中比特1所占整个序列的比例。理论上，随机序列的0游程和1游程应是相等的。检测过程中，若 r_i 为游程长度为 i 时的0游程或1游程的出现频数，而理论游程频数与序列长度间的关系满足[19]：

$$\frac{n}{8}=r_1=2r_2=2^2r_3=\cdots=2^{i-1}r_i=\cdots\tag{3}$$

游程检测需构造服从标准正态分布函数的统计量 $\frac{V_n-2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)}$，对其服从标准正态分布函数作如下分析：

首先，游程检测建立在码元频数检测基础上，计算获得 π 后，即可得二项式和 $S_n=n\pi-n(1-\pi)$，而 $\frac{S_n}{\sqrt{n}}=\frac{n\pi-n(1-\pi)}{\sqrt{n}}$ 接近于标准的正态分布函数，即满足 $\lim\limits_{n\to\infty}P(\frac{n\pi-n(1-\pi)}{\sqrt{n}}\leqslant z)=\phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-u^2/2}du$。同时，当0或1游程的游程长度值最大时该游程出现的频数近似等于1，根据理论游程频数与序列长度间的关系可得不同长度的游程频数满足：

$$\begin{cases}r_1=\dfrac{n}{8}\\r_2=\dfrac{n}{16}\\r_3=\dfrac{n}{32}\\\cdots\\r_{i_{\max}}=\dfrac{n}{2^{i_{\max}+2}}\end{cases}\tag{4}$$

因此，0游程或1游程的游程总数可表示为 $Sum_r=\frac{n}{4}-\frac{n}{2^{i_{\max}+2}}$，$i_{\max}$ 为最长游程的长度，而理论游程总数则满足 $Sum_{idea}=2Sum_r=\frac{n}{2}-\frac{n}{2^{i_{\max}+1}}$。最长游程长度 $i_{\max}$、游程频数 $r_{i_{\max}}$ 及序列长度 n 间应满足 $\frac{n}{8}=2^{i_{\max}-1}r_{i_{\max}}$，其中 $r_{i_{\max}}=1$ 或 $r_{i_{\max}}=2$，理论游程总数 $V_{nidea}=Sum_{idea}=2Sum_r=\frac{n}{2}-\frac{n}{2^{i_{\max}+1}}\approx\frac{n}{2}-2\overset{n\to\infty}{\approx}\frac{n}{2}$。而 $\frac{n\pi-n(1-\pi)}{\sqrt{n}}$ 具有如下形式。

$$\frac{n\pi-n(1-\pi)}{\sqrt{n}}=\frac{2\pi(1-\pi)[n\pi-n(1-\pi)]}{2\sqrt{n}\pi(1-\pi)}$$

$$=\frac{4n\pi^2(1-\pi)-2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)}$$

$$\approx\frac{V_n-2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)} \tag{5}$$

综上所述，$Run_n=\frac{V_n-2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)}$服从标准的正态分布函数。游程检测定义统计门限值 P-value 为：

$$P\text{-value}=2[1-\varphi(Run_n)]=erfc(\frac{Run_n}{\sqrt{2}})$$

$$=erfc(\frac{V_n-2n\pi(1-\pi)}{2\sqrt{2n}\pi(1-\pi)}) \tag{6}$$

2.3 块内频数检测

块内频数检测主要用于检测长度为 M 的比特分块内，比特 0 和 1 的出现频率是否符合随机性要求。块内频数检测以 M 为分块长度，采用非重叠分块模式将长度为 n 的比特序列划分为 $N=\lfloor\frac{n}{M}\rfloor$ 块待检测比特块，并计算每一分块内比特 1 出现的比例 $\pi_i=\frac{\sum_{j=1}^{M}\varepsilon_{(i-1)M+j}}{M}$，$1\leqslant i\leqslant N$。在确定各分块内比特 1 出现的概率后，构造 $\chi^2(obs)=4M\sum_{i=1}^{N}(\pi_i-\frac{1}{2})^2$，而 $\chi^2(obs)$ 服从自由度为 $N/2$ 的 χ^2 分布，统计门限值 P-value 则被定义为：

$$P\text{-value}=igamc(N/2,\chi^2(obs)/2)$$

$$=\frac{1}{\Gamma(N/2)}\int_{\chi^2(obs)/2}^{\infty}e^{-t}t^{N/2-1}dt \tag{7}$$

其中，$\Gamma(N/2)=\int_0^{\infty}t^{N/2-1}e^{-t}dt$。

2.4 存在的问题

分析 SP800-22 rev1a 标准中码元频数、游程及块内频数检测的原理及实现过程不难发现，影响码元频数、游程检测精度的主要参数为待检测比特序列的长度，而比特序列长度和分块长度则是影响块内频数检测精度的主要参数。下面将分析以上参数对检测精度的影响。

(1)比特序列长度对检查精度的影响

SP800-22 rev1a 标准建议待检测比特序列的最短长度不应小于 100 比特。以某无线网络链路层数据的比特序列为分析对象，分别测试了码元频数、游程、块内频数检测(分块长度为 30 比特)的识别能力。在比特序列的平均长度为 150 比特时，未加密比特序列的识别率分别为 1.3%、2.1%、1.1%，而已加密比特序列的识别率则分别为 99.5%、99%、98.5%。在比特序列长度为 600 比特时，未加密比特序列的识别率分别为 6.4%、25.9%、10.3%，而已加密比特序列的识别率为 99.1%、98.8%、98.6%。当比特序列长度较短时，三者均无法有效对未加密比特序列进行识别。

(2)分块长度对识别率的影响

SP800-22 rev1a 标准建议的比特序列的分块长度最短不应小于 20 比特，但并未给出具体的分块长度选择方法。在比特序列的平均长度为 300 比特条件下，对块内频数检测的识别率进行了测试。在采用块内频数检测方法对已加密比特序列进行识别时，分块长度对其识别率不会产生较大的影响，平均识别率可达 98.8%。对未加密的比特序列而言，分块长度对识别率的影响则较大，在分块长度为 20～60 比特的条件下，识别率变化范围为 0.3%～8.2%。

以上测试结果表明：在未对以上检测方法进行参数优化前，以链路层比特序列中比特 0 和 1 的随机分布程度作为判断数据是否加密的主要依据时，上述随机性检测方法对已加密的比特序列的识别率较高，而对未加密比特序列的识别率则很低。

3 链路层加密数据识别方案

根据以上测试方法的检测能力及 SP800-22 rev1a 标准中建议的检测顺序，首先对链路层数据的比特序列进行码元频数检测，比特序列的长度较长时，采用码元频数检测可取得较为理想的结果。对于长度较短的比特序列，多数情况下会将未加密比特序列判定为已加密比特序列。对于码元频数检测阶段被判断为可能已加密的比特序列，进一步采用游程检测方法检测比特序列内比特 0 和 1 的变化频率是否符合随机性检测的要求。码元频数检测及游程检测均无法有效反映比特序列的中 0 和 1 的局部分布特征，而块内频数检测则用于检测比特分块中 0 和 1 的分布是否满足随机性要求。从链路层数据的比特序列构成特点来看，块内频数检测能够更好地挖掘比特序列中比特 0 和 1 的局部分布特征，但仍需对其参数进行优化，以提高对未加密比特序列的识别率。

3.1 分块长度候选值选择

对于已加密的链路层比特序列，其整个序列内比特 0 和 1 的出现概率接近于 50%。从块内频数检测的本质来看，在分块长度不小于 20 比特时，各分块内比特 0 和 1 的出现概率也应接近于 50%。基于方差原理的分块长度候选值选择方案的主要过程如下：

Step 0：令分块长度 $M=20$，R 条待识别比特序列中长度最短的比特序列的长度为 $L_{\min}$，限定分块长度最大值 $M_{\max}=\lfloor\frac{L_{\min}}{2}\rfloor$。

Step 1　若 $M\leqslant M_{\max}$，以 M 为分块长度对 R 条序列进行分块得 $BLOCK_{1M}$，$BLOCK_{2M}$，$BLOCK_{3M}$，…，$BLOCK_{RM}$，其中 $BLOCK_{iM}=(block_{i1},block_{i2},block_{i3},\cdots,block_{i\lfloor\frac{Len_i}{M}\rfloor})$。

Step 2　$BLOCK_{1M}$，$BLOCK_{2M}$，$BLOCK_{3M}$，…，$BLOCK_{RM}$ 中比特 1 出现的概率为 $\pi_{1M},\pi_{2M},\pi_{3M},\cdots,\pi_{RM}$，其中 $\pi_{iM}=(pi_{i1},pi_{i2},pi_{i3},\cdots,pi_{i\lfloor\frac{Len_i}{M}\rfloor})$。

Step 3　统计 $X=(x_1,x_2,x_3,\cdots,x_{M+1})=(0,\frac{1}{M},\frac{2}{M},\cdots 1)$中的元素在 $\pi_{1M},\pi_{2M},\pi_{3M},\cdots,\pi_{RM}$ 中的出现概率 $P_1,P_2,P_3,\cdots,P_R$，而 $P_i=(p_{i1},p_{i2},p_{i3},\cdots,p_{i(M+1)})$。

Step 4　统计 $P_1,P_2,P_3,\cdots,P_R$ 的平均方差 $\tilde{D}_M(P)=\frac{\sum_{i=1}^{R}\sum_{j=1}^{M+1}p_{ij}(p_{ij}-\tilde{P}_i)^2}{R}$，并令 $M=M+1$。

Step 5　重复执行 Step 1－Step 4，获得统计量 $P_1,P_2,P_3,\cdots,P_R$ 对于 R 条比特序列的方差均值$\tilde{D}_{20}(P)$，$\tilde{D}_{21}(P)$，…，$\tilde{D}_{M_{\max}}(P)$。

Step 6　根据等式 $\Delta D_i(P)=\tilde{D}_{i+20}(P)-\tilde{D}_{i+19}(P)$计算 $\tilde{D}_{20}(P)$，$\tilde{D}_{21}(P)$，…，$\tilde{D}_{M_{\max}}(P)$中相邻元素的差值 $\Delta D_1(P)$，$\Delta D_2(P)$，…，$\Delta D_{M_{\max}-20}(P)$。若 $\Delta D_i(P)\geqslant\theta$ 则 $i+20$ 即为分块长度候选值，其中 θ 为确定分块长度候选值的阈值。

提出的方案通过度量 $P_i=(p_{i1},p_{i2},\cdots,p_{i(M+1)})$相对于其

期望值 $\tilde{P}_i$ 的偏移程度 $D(P_i)=\sum_{j=1}^{M+1} p_{ij}(p_{ij}-\tilde{P}_i)^2$ 以确定可能的分块长度候选值，即要求分块长度候选取值使 $D(P_i)$ 尽可能大且增量明显。满足此类条件的分块长度值，可使得统计量 $P_i=(p_{i1},p_{i2},\cdots,p_{i(M+1)})$ 中的值分布更为分散，进而能有效提高对未加密比特序列的识别率。

3.2 分块长度目标值确定

多次随机实验结果表明，在众多的分块长度候选值中，分块长度目标值均包含于候选值集合内，且目标值处的方差增量较大，方差变化量较小。基于方差变化量的分块长度目标值确定方案如下：

Step 0 根据需求设定 $Repeat_Num$，$Count=1$，R_1。

Step 1 当 $Count\leqslant Repeat_Num$ 时，采用随机方式从 R 条待识别比特序列中选择 R_1 条对其执行分块长度候选值选择方案，获得对应的分块长度候选值集合 $A_{Count}=\begin{bmatrix} mA_{Count} \\ \Delta D \end{bmatrix}=\begin{bmatrix} m_1 & m_2 & \cdots & m_j \\ \Delta D_{m_1-20}(P) & \Delta D_{m_2-20}(P) & \cdots & \Delta D_{m_j-20}(P) \end{bmatrix}$，并令 $Count=Count+1$。

Step 2 重复执行步骤 1，获得统计量 $(A_1,A_2,\cdots,A_{Repeat_Num})$，并求解 $mA_1\cdots\bigcap mA_{Repeat_Num}$，以缩小分块长度候选值范围。

Step 3 计算在候选长度为 m_i 时集合 $(A_1,A_2,\cdots,A_{Repeat_Num})$ 中 ΔD_{m_i} 的均值 $\Delta\tilde{D}_{m_i}$ 及其方差 $D(\Delta D_{m_i})$。根据 $R_{D_{mi}}=\alpha\Delta\tilde{D}_{m_i}-\beta D(\Delta D_{m_i})$ 计算 $R_{D_{mi}}$。

Step 4 选择使 $R_{D_{mi}}$ 取得最大值的分块长度作为分块长度目标值。

$R_{D_{mi}}=\alpha\Delta\tilde{D}_{m_i}-\beta D(\Delta D_{m_i})$ 中的参数 α 和 β 用于表征方差的变化量及其方差对判定值的影响。后续实验取 $\alpha=\beta=0.5$。

3.3 比特序列信息提取

为提高对长度较短的未加密比特序列的识别率，将随机抽样原理引入到比特序列信息的提取过程中。基于随机抽样原理的比特序列信息提取过程如图 1 所示。

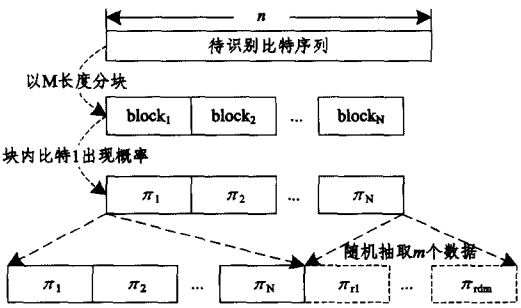

图 1 比特序列信息提取

m 次随机抽样后，$\chi^2(obs)=4M\sum_{i=1}^{N}(\pi_i-1/2)^2$ 更新为 $\chi^2_{New}(obs)=4M[\sum_{i=1}^{N}(\pi_i-1/2)^2+\sum_{i=1}^{m}(\pi_{ri}-1/2)^2]$，且 $\chi^2_{New}(obs)$ 服从自由度为 $(N+m)/2$ 的 χ^2 分布。比特序列已加密时，任意分块 $block_i$ 内比特 1 出现的概率 π_i 应以不可忽略的概率满足 $|\pi_i-1/2|\leqslant\xi$，而 $\Delta\chi^2(obs)_1=\sum_{i=1}^{N_1}(\pi_{ri}-1/2)^2$ 亦应以不可忽略的概率满足 $\Delta\chi^2(obs)_1=\sum_{i=1}^{m}(\pi_{ri}-1/2)^2\leqslant m\xi^2$。未加密时，$\pi_i$ 则以不可忽略的概率满足 $|\pi_i-1/2|\geqslant\xi$，而 $\Delta\chi^2(obs)_2=\sum_{i=1}^{m}(\pi_{ri}-1/2)^2$ 亦应以不可忽略的概率满足 $\Delta\chi^2(obs)_2=\sum_{i=1}^{m}(\pi_{ri}-1/2)^2\geqslant m\xi^2$。在相同自由度条件下，有 $\Delta P\text{-value}_1\leqslant\Delta P\text{-value}_2$ 成立，可见随机抽样对未加密比特序列的统计门限值 P-value 的影响更大。图 2 为不同随机抽样次数下，满足 $\Delta\chi^2(obs)_1<\Delta\chi^2(obs)_2$ 的概率统计值。

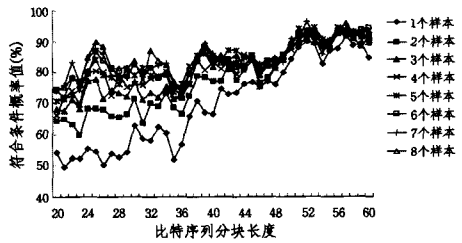

图 2 $\Delta\chi^2(obs)_1<\Delta\chi^2(obs)_2$ 的概率统计值

4 实验结果与分析

4.1 分块长度候选值选择实验

在选择分块长度候选值的过程中，$P=(p_0,p_1,p_2,\cdots,p_M)$ 的方差随分块长度变化的规律如图 3 所示，$P=(p_0,p_1,p_2,\cdots,p_M)$ 的方差随分块长度的增加呈递增趋势，当分块长度大于某值时方差趋于稳定。如图 4 所示，以 $\theta_1=0.1\tilde{D}_{\max}$ 为分块长度候选值筛选阈值，筛选出 30、37、43、49、52 为分块长度候选值。

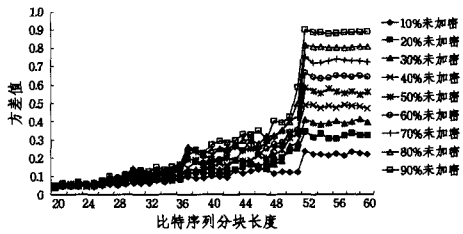

图 3 不同分块长度条件下方差变化情况

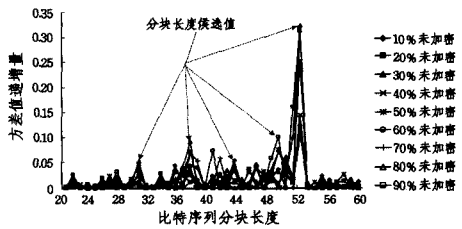

图 4 比特序列分块长度候选值分布情况

4.2 分块长度目标值分析与确定

基于方差原理的分块长度候选值选择过程确定了 30、37、43、49、52 作为分块长度候选值。由 $R_{D_{mi}}=0.5\Delta\tilde{D}_{m_i}-0.5D(\Delta D_{m_i})$ 计算出的分块长度判定值如表 1 所列。根据分块长度目标值确定原则，应选择 52 作为块内频数检测过程中的比特序列分块长度目标值。

表 1 分块长度判定值

分块长度	30	37	43	49	52
方差变化量均值	0.0270	0.0440	0.0250	0.0450	0.2230
变化量的方差值	0.0003	0.0007	0.0004	0.0448	0.0086
判定值	0.0134	0.0217	0.0123	0.0001	0.1072

4.3 随机抽样对识别率的影响实验

在比特序列的平均长度为150比特，分块长度为52比特的条件下，进一步验证了对$(\pi_1,\pi_2,\pi_3,\cdots,\pi_N)$进行随机抽样给识别率造成的影响，结果如图5所示。未随机抽样时，未加密比特序列的平均识别率仅为7.5%，已加密比特序列的平均识别率为99.8%。在随机抽样次数为20的条件下，未加密比特序列的平均识别率可提高到85.2%，而已加密比特序列的平均识别率则降低到91.6%，其主要原因在于随机抽样次数的增加降低了序列$(\pi_1,\pi_2,\pi_3,\cdots,\pi_N)$所代表比特序列的随机程度。在满足被检测报文序列均具有相似属性及样本数量超过20的条件下，文献[7]提出算法对加密流量达到了90%以上的识别率，在样本数量为10时，其识别率达到了80%以上。而本文提出的方案对1条未加密数据的识别率即为85.2%，对已加密数据的识别率为91.6%，且提出方案的识别率不存在报文间的干扰问题，也无需对报文序列进行预处理，在保证识别率的同时具有较好的适用性。

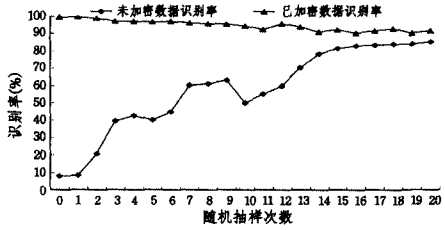

图5 随机抽样次数对识别率的影响

4.4 识别出的未加密比特序列的可信度分析

以上从实验角度对方案的可行性进行了说明，下面将从概率论的角度对块内频率检测识别出的未加密比特序列的可信度进行分析。为便于描述，定义事件A={被检测比特序列为未加密比特序列}、B={随机性检测结果P-value<0.01}，正确检测未加密比特序列的概率为$P(B|A)$，已加密比特序列检测错误的概率为$P(B|\overline{A})$，则在P-value<0.01条件下，比特序列为未加密比特的概率可表示为：

$$P(A|B)=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\overline{A})P(B|\overline{A})} \tag{8}$$

在概率公式$P(A|B)$中，一般$P(A)$为未加密比特序列在比特序列中所占比例，令其为$r\in[0,1]$，令$P(A|B)=z$、$P(B|A)=x$、$P(B|\overline{A})=y$，则公式$P(A|B)$可表示为：

$$z=\frac{rx}{rx+(1-r)y} \tag{9}$$

如前所述，在随机抽样次数为20的条件下，未加密比特的平均识别率为85.2%(即$P(B|A)=0.852$)，而已加密比特序列的平均识别率则为91.6%($P(B|\overline{A})=0.084$)。图6为不同$P(A)$取值条件下，$P(A|B)$取值的统计结果。如图6所示，同一般的估计方法相比，在未加密比特序列所占比例较小情况下，采用随机性检测方法，可显著提高未加密比特序列识别结果的可信度。鉴于分析已加密比特序列识别结果的可信度，可采用上述相同的方法，不再赘述。

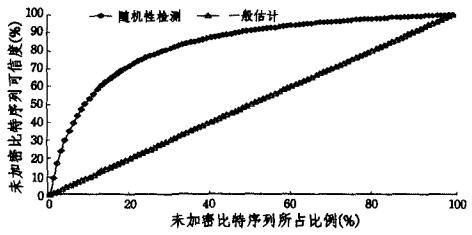

图6 $P(A|B)$统计结果

结束语 链路层加密数据识别对加强网络安全防护能力及进一步的协议识别技术研究具有重大意义，从链路层未加密数据与已加密数据的随机统计特性角度出发，提出了一种链路层加密数据识别方案。通过在块内频数检测过程中引入方差原理及随机抽样原理，解决了分块长度选择及未加密数据识别率低的问题。理论分析及实验结果均表明，提出的方案在一定程度上解决了链路层加密数据识别问题且具有较高的识别率。

参 考 文 献

[1] 陈亮，龚俭，徐选. 应用层协议识别算法综述[J]. 计算机科学，2007，34(7)：73-75

[2] 吴志军，阚洪涛. 基于ECC的TES网络链路层安全协议的研究[J]. 通信学报，2009，30(11)：86-92

[3] 张杨. 基于NDIS的数据链路层加密[J]. 辽宁大学学报，2010，37(2)：116-118

[4] DOD(Department of Defense). MIL-STD-188-220C[EB/OL]. http://www.everyspec.com，2012-06-29

[5] Charles V W，Fabian M，Gerald M M. On inferring application protocol behaviors in encrypted network traffic[J]. Journal of Machine Learning Research，2006，7(12)：2745-2769

[6] Sun Guang-lu，Xue Yi-bo，Dong Ying-fei，et al. A Novel Hybrid Method for Effectively Classifying Encrypted Traffic[C]// Proceedings of Communications and Systems Security，2010(GLOBECOM 2010). Miami USA：IEEE，2010：1-5

[7] 赵博，郭虹，刘勤让，等. 基于加权累积和检验的加密流量盲识别算法[J]. 软件学报，2013，24(6)：1334-1345

[8] Menezes Alfred J，van Oorschot P C，Vanstone S A. 应用密码学手册[M]. 胡磊，王鹏，等译. 北京：电子工业出版社，2005

[9] 胡俭勇，苏锦海. 一种随机性实时检测方案[J]. 计算机工程，2009，34(9)：136-138

[10] Rukhin A，Soto J，Nechvatal J. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications[R]. Technical Report. SP 2001：800-22

[11] Filiol E. A new statistical testing for symmetric ciphers and hash functions[C]// Information and Communications Security：4th International Conference. Berlin：Springer，2002：342-353

[12] NIST FIPS PUB 140-2-2001. Security Requirements for Cryptographic Modules[S]. Washington DC，USA：National Institute of Standards and Technology，http://csrc.nist.gov，2001

[13] NIST SP800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications Revision 1a[S]. Washington DC，USA：National Institute of Standards and Technology，http://csrc.nist.gov，2010