计算机科学
COMPUTER SCIENCE

基于信息熵与闭合频繁序列的密码协议逆向方法

梁晨, 洪征, 吴礼发, 吉庆兵

引用本文

梁晨, 洪征, 吴礼发, 吉庆兵. 基于信息熵与闭合频繁序列的密码协议逆向方法[J]. 计算机科学, 2024, 51(3): 326-334.

LIANG Chen, HONG Zheng, WU Lifa, JI Qingbing. Cryptographic Protocol Reverse Method Based on Information Entropy and Closed Frequent Sequences [J]. Computer Science, 2024, 51(3): 326-334.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200147

基于信息熵与闭合频繁序列的密码协议逆向方法

梁　晨[1]　洪　征[2]　吴礼发[1]　吉庆兵[3]
1 南京邮电大学网络空间安全学院　南京 210023
2 陆军工程大学指挥控制工程学院　南京 210007
3 中国电子科技集团公司第三十研究所　成都 610041
(1020041310@njupt.edu.cn)

摘　要　未知密码协议被广泛用于敏感信息的安全传输,对其进行逆向分析对攻防双方都具有重要意义。为从网络流量中推断结构复杂的密码协议格式,提出了一种基于信息熵与闭合频繁序列的密码协议逆向方法。利用字节信息熵划分报文的明文域与密文域,使用BIDE算法挖掘闭合频繁序列,划分报文的动态域和静态域;设计了一种长度域识别算法,对报文进行字节片切,将片切后的字段值与长度域取值集合进行循环比对,实现了密码协议中多种形式的长度域识别;设计了启发策略,用于对加密套件、加密算法等密码协议特有的关键字段进行语义识别。实验结果表明,该方法可以有效地对密码协议进行域划分,提取密码协议的格式,并且在长度域识别和密码协议特有关键字段的语义识别上优于现有方法。

关键词:协议逆向;密码协议;信息熵;闭合频繁序列;网络流量;语义分析
中图分类号　TP393

Cryptographic Protocol Reverse Method Based on Information Entropy and Closed Frequent Sequences

LIANG Chen[1], HONG Zheng[2], WU Lifa[1] and JI Qingbing[3]
1 School of Cybersecurity, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
2 College of Command and Control Engineering, Army Engineering University, Nanjing 210007, China
3 No. 30 Institute of CETC, Chengdu 610041, China

Abstract　Unknown cryptographic protocols are widely used for the secure transmission of sensitive information, and reversing cryptographic protocol is of great significance to both attackers and defenders. In order to efficiently reverse complex cryptographic protocols, a cryptographic protocol reverse method based on information entropy and closed frequent sequences is proposed. The information entropy is used to distinguish the plaintext and ciphertext, and the closed frequent sequences mined by BIDE algorithm are used to identify dynamic fields and static fields in the messages. A length field identification algorithm is proposed. It slices the message, and compares the sliced field values with the set of length field values to achieve various forms of length field recognition in cryptographic protocols. Heuristic strategies are proposed to recognize the semantics of key fields including the fields specific to cryptographic protocols such as encryption suites and encryption algorithms. Experimental results show that the method can effectively identity fields and extract the formats of cryptographic protocols, outperforms the existing methods in various length fields identification and semantic recognition of key fields specific to cryptographic protocols as well.

Keywords　Protocol reverse, Cryptographic protocol, Information entropy, Closed frequent sequence, Network traffic, Semantic recognition

1　引言

近年来,随着网络安全保护意识的提升和密码技术的发展,用于安全保密通信的标准密码协议得到了广泛应用,如IPsec,PGP,HTTPS,FTPS等,它们使用数据加密、数字签名、完整性校验等各种技术对用户通信进行保护。此外,军事通信系统和很多恶意代码会采用未知的密码协议进行保密通信,对这些私有的密码协议进行逆向分析对攻防双方而言都具有重要意义。

协议逆向分析方法主要分为两类:基于网络流量的分析方法和基于执行轨迹的分析方法。前者通过捕获协议的网络数据包,来分析字节流的取值变化率以进行特征推断,进而

到稿日期:2022-12-25　返修日期:2023-04-06
基金项目:国家重点研发计划(2019YFB2101704)
This work was supported by the National Key Research and Development Program of China(2019YFB2101704).
通信作者:吴礼发(wulifa@njupt.edu.cn)

得到协议格式；后者利用动态分析技术，跟踪程序对协议数据包的解析过程，分析程序执行轨迹来提取协议格式。基于网络流量的分析方法主要分为 4 类：基于序列比对的格式推断算法、基于概率模型的格式推断算法、基于频繁项集的格式提取算法和基于语义的格式推断算法[1]。文献[1-4]对近年来基于网络流量的协议逆向分析领域的典型方法、工具或系统进行了分析与总结。但以上方法主要针对明文协议，无法直接应用在结构更为复杂的密码协议中。

对于密码协议，文献[5-9]利用动态污点分析技术分析程序执行轨迹，实现对加密报文的解析和格式提取，局限性是需要获得协议通信软件，且涉及复杂的二进制程序分析技术。Zhu 等[10]提出了 SPFPA 方法，利用信息熵推断加密字段并且提出了加密字段的边界定位算法。He[11]利用网络轨迹并基于贪心算法实现了对密码协议的解析。Tang 等[12]利用滑动窗口方法计算分段熵，对加密报文进行特征提取。

基于网络流量进行密码协议逆向分析的主要难点在于如何对存在密文域的报文进行准确高效的报文结构推断，识别形式多样的长度域字段。目前针对密码协议的逆向分析方法是对基于频繁模式的明文协议的逆向方法进行改进，但存在以下问题：首先，密码协议存在复杂的变长字段，挖掘频繁模式对关键字段的位置要求严格，直接应用频繁模式挖掘算法进行域划分的效率较低；其次，在结构推断方面存在长度域关键词识别不完善等问题，无法识别一些位置特殊、与指示字段不相邻、指示字段非密文域的长度域关键词；最后，对关键字段的语义识别不够完善，缺少一些针对密码协议特有字段的识别策略。

针对上述问题，本文设计了一种基于信息熵和闭合频繁序列的密码协议逆向分析方法(a Reverse Approach of security protocol based Information Entropy and Closed Frequent Sequences，RAECFS)。该方法以密码协议网络流量为输入，采用信息熵和闭合频繁序列对报文域进行划分，利用设计的长度域识别算法实现多长度域识别，最后利用报文间的相似性与差异性和关键字段特点推断其语义。综上所述，本文的主要贡献有：

1)设计了一种密码协议的域划分方法，计算信息熵定位密文域，将密文域进行替换后生成两方向报文组，分别挖掘闭合频繁序列，划分动态域和静态域，并辅助关键字段的语义识别工作。

2)总结分析了密码协议的长度域特点，提出了一种密码协议长度域识别算法，对报文进行片切，将片切后的字段值与长度取值集合进行循环比对，实现密码协议中多种形式的长度域识别。

3)总结分析了密码协议关键字段的特点，针对校验和、标识符等常见字段和加密算法、加密套件等密码协议特有的关键字段，设计了启发策略进行语义识别。

2　密码协议特点分析

目前，互联网中常见的密码协议有安全 IP 协议(IPsec)、安全套接层协议(SSL/TLS)、安全外壳层协议(SSH)、消息流加密协议(MSE)、计算机网络认证协议(Kerberos)和局域网扩展认证协议(EOPAL)等。密码协议一般由握手协议和传输协议两部分组成。在密码协议的握手阶段，通信双方要进行建立连接、认证身份、协商加密算法等工作。在密码协议的传输阶段，报文需要传输加密载荷、消息摘要等密文数据。下面对密码协议报文的格式进行分析。

1)报文结构更为复杂。报文内部的关键字段具有顺序、并列和递进关系。密码协议的安全性不仅依赖于所用的密码算法的安全强度，还与协议的报文结构有密切关系[13]。图 1 所示报文中的 KEX host key 与 KEX H signature 是顺序关系，两者共同组成了 Key Exchange 字段。

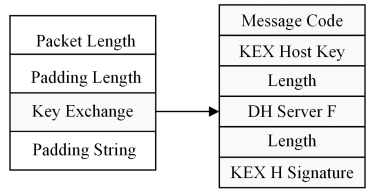

图 1　SSH 协议 Diffie-Hellman Key Exchange Reply 报文

Fig. 1　Diffie-Hellman Key Exchange Reply message in SSH protocol

2)存在多种类型的密文域。密码协议需要使用加密算法、数字签名、哈希算法等多种密码技术对关键信息进行加密，密文域可能是变长或定长的随机数、密钥、认证码等加密数据。表 1 列出了密码协议中几类常见的密文域的构成形式。

表 1　常见密文域分布形式

Table 1　Common kinds of ciphertext format

类别	形式
定长	[Keyword][Ciphertext]
	[Keyword][Plaintext][Ciphertext]
	[Keyword][Length][Ciphertext]
	[Keyword][Length][Plaintext][Ciphertext]
变长	[Keyword][Ciphertext]
	[Keyword][Length][Ciphertext]

其中，Keyword 代表静态关键字段，Ciphertext 代表密文域，Plaintext 代表明文域，Length 代表长度域字段。

3)长度域的形式复杂。密码协议存在分别指示不同字段的多个长度域，且位置不固定。

长度域字段一般位于指示字段之前，但不一定相邻。图 2 所示的报文存在两个长度域且与其指示的字段不相邻。本文总结了如表 2 所列的长度域和其指示字段的存在形式与结构。

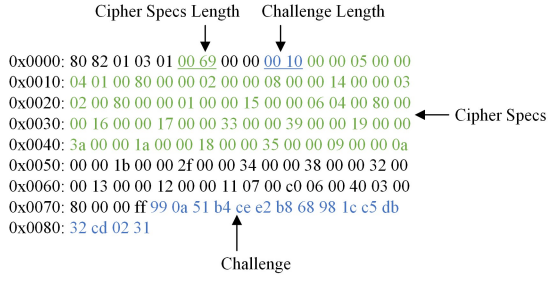

图 2　SSL 协议的 ClientHello 报文

Fig. 2　ClientHello message in SSL protocol

表 2 长度域字段的存在形式

Table 2 Common types of length field forms

序号	形式
1	$Length_i[CipherText]_i Length_i[CipherText]_i$
2	$Length_i[[Plaintext][CipherText]]_i$
3	$Length_i[Length_i[PlainText]_j\ Length_k[CipherText]_k]_j$
4	$Length_j[PlainText]\ [PaddingString]_i$
5	$Length_i[CipherText]_i\ Length_j[CipherText]_j\ Length_k[CipherText]_k$

表 2 中，Length 代表长度域，其所标识的作用域分为 3 种类型，即明文域 PlainText、填充域 PaddingString 和密文域 Ciphertext。下标用于关联长度域字段和其指示的具体字段。

3 逆向方法的总体框架

本文综合明文协议的逆向方法和密码协议报文特征，设计了一种密码协议逆向方法，如图 3 所示，其具体分为 3 个阶段。

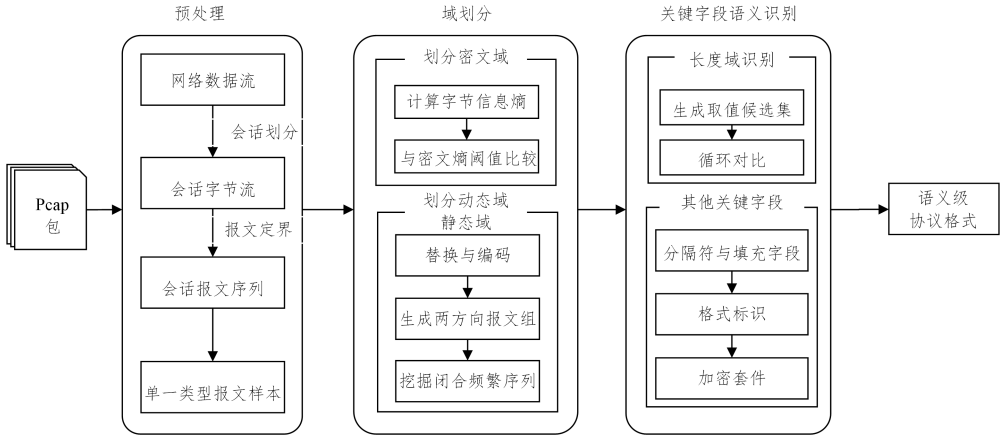

图 3 RAECFS 方法的框架

Fig. 3 Framework of RAECFS

4 密码协议域划分方法

密码协议由密文域和明文域组成，明文域可根据字节取值是否可变划分为静态域和动态域。本文设计了一种如图 4 所示的密码协议域划分方法。

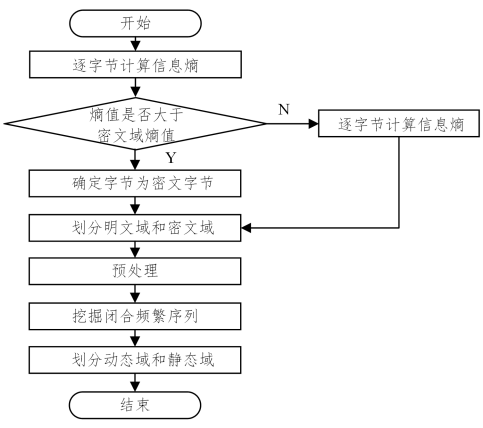

图 4 域划分流程

Fig. 4 Process of field identification

1)密文域划分。获得纯净的单一协议报文序列后，逐字节计算其报文垂直方向的信息熵，根据熵值判断明密文属性，合并同属性的相邻字节，划分密文域和明文域。

2)静态域与动态域划分。对报文序列进行预处理后，生成垂直报文组和水平报文组，挖掘闭合频繁序列，将明文域划分为动态域和静态域。

4.1 密文域划分

本阶段利用信息熵划分密码协议的明文域和密文域。

信息熵是对事物运动状态或存在方式的不确定性的描述，不确定性越大熵值越大[14]，可用来反映字节取值的随机性程度和协议序列中字节的分布特性。密文数据加密后失去统计特征，近似为随机取值。

文献[15]指出静态域中的字节熵值接近 0，密文域熵值较大，动态域熵值处于两者之间。本文通过实验对此进行了验证。

定义报文中偏移位置 i 的字节的垂直方向信息熵为(报文总数为 N)：

$$H(x_i)=-\sum_{k=1}^{255} f_k \log_2 f_k \tag{1}$$

其中，i 是该字节在协议流量序列中的序号位置(即偏移量)，f_k 表示在 N 个报文中所有第 i 个字节取值为 k 的频率。

1)数据预处理。获得原始流量后进行分组过滤、删去重传，得到协议的流量样本。将流量样本按照时间阈值构建为不同会话，按照会话中的报文偏移得到单一种类的报文序列。

2)域划分。域划分包括两个步骤，首先是密文域和明文域的划分：逐字节计算报文序列的信息熵，与密文信息熵阈值进行比较，判断该字节的明密文属性，合并连续偏移的同属性字节。其次，对报文序列的动态域和静态域进行划分：对报文预处理后按时序关系和会话逻辑关系生成垂直报文组与水平报文组，挖掘闭合频繁序列，划分动态域和静态域。本文将在第 4 节介绍域划分方法。

3)关键字段语义识别。在域划分结果的基础上，对密码协议的关键字段进行语义识别。生成长度域可能取值的候选集，进行报文字节片切和循环比对以识别多个长度域字段；根据密码协议中关键字段的特点设计不同的启发策略，对其他关键字段进行语义识别。本文将在第 5 节详细讨论关键字段语义识别方法。

计算 SSH 协议的 Diffie-Hellman Group Exchange 报文前 140 个字节的信息熵，结果如图 5 所示，其中 11～140 字节的密文域熵值在 7.05～7.20 之之间，与明文域存在显著差异。

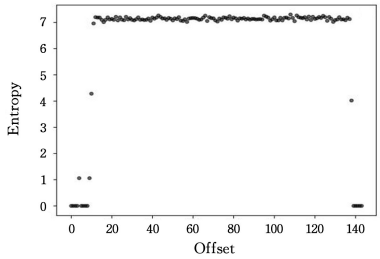

图 5　Diffie-Hellman Group Exchange 报文的信息熵

Fig. 5　Information entropy of each byte in Diffie-Hellman Group Exchange messages

计算 SSH 协议的 Encrypted packet 报文前 54 个字节的信息熵，结果如图 6 所示。熵值均分布在 7.05～7.20 之间，但长度域、数据字段、消息认证码之间的熵值没有显著差异。

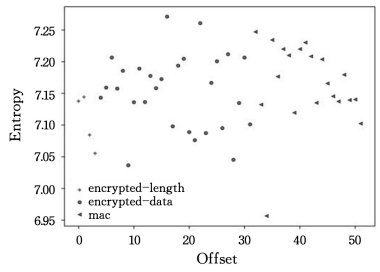

图 6　Encrypted packet 报文的信息熵

Fig. 6　Information entropy of Encrypted packet messages

以上结果显示，明文字节和密文字节的熵值存在显著差异，密文字节中的不同关键字段的熵值近似相同。因此，选择信息熵作为明文字节和密文字节的区别特征。

密文域中的字节变量取值近似满足均匀分布[16]。当样本数量无穷多时，取值随机的密文字节的信息熵为 8，实际环境下无法获得无穷多样本，因此选择 N-截断熵作为阈值判断字节明密文属性。截断熵 $H_N(p)$ 指样本数量为 N 时，字节依据概率分布 p 时的简单熵 $\hat{H}_N^{\mathrm{MLE}}(x)$ 的平均值。随机生成 N 个随机取值的字节，计算字节的信息熵得到平均值，$H_N(p)$ 与样本数量的关系如图 7 所示。可通过最大似然估计计算 N-截断熵[17]，在信息源满足均匀分布时，N-截断熵为无偏估计值。

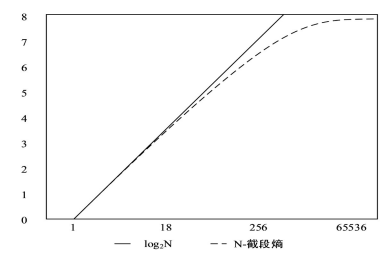

图 7　$H_N(p)$ 曲线

Fig. 7　$H_N(p)$ curve

逐字节计算信息熵，比较熵值与 N-截断熵，若误差小于标准误差则该字节为密文字节，若大于标准误差则该字节为明文字节。获得字节属性后，将连续偏移的相同属性字节合并，完成明文域和密文域的划分。

4.2　动态域和静态域划分

接下来对明文域做进一步划分。本节先对相关概念给出定义与解释，然后对划分过程进行详细描述。

4.2.1　支持度与闭合频繁序列

支持度指子序列在报文样本中出现的次数占总样本数量的比例。

频繁序列指在报文中出现频率大于支持度阈值的字节序列。Apriori 性质[18]指出，一个非频繁序列的任意超序列也是非频繁项序列。若某频繁序列不存在支持度大于该序列支持度的超序列，则其为报文组的闭合频繁序列。

4.2.2　报文字节序列处理

为了提高闭合频繁序列的挖掘效率并降低计算开销，对报文进行了处理，具体如图 8 所示。

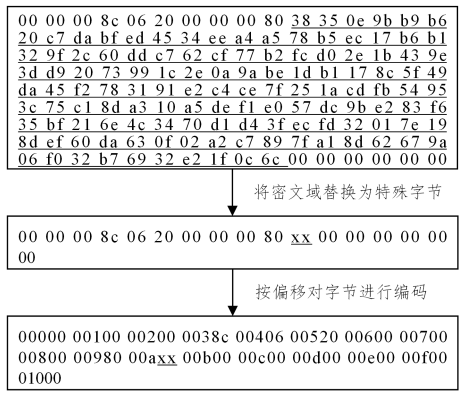

图 8　报文字节序列的处理过程

Fig. 8　Message processing process

1）替换：将先前划分的密文域替换为一个特殊字符串“xx”，不参与动态域和静态域划分，同时保证其后的关键字段相对位置不改变。

2）编码：为了将字节与偏移位置相关联，我们区分了不同偏移位置下的相同字节序列，将字节序列按照偏移位置进行编码。在字节前增加 3 位 16 进制数表示字节的偏移，即 $Byte\_token=(Byte.loc, Byte.val)$。

4.2.3　报文组生成

不同阶段的报文结构相似，利用报文间的关联性，生成垂直方向的报文组和水平方向的报文组分别进行闭合频繁序列挖掘。

本文方法的分析样本是单报文流会话，记为 $S=\{m_1, m_2, \cdots, m_N\}$，$N$ 为报文总数。会话集合记为 $P=\{s_1, s_2, \cdots, s_x\}$，$x$ 是分析样本中的会话总数。根据报文的时序关系和会话逻辑关系，生成两方向报文组，具体方法如图 9 所示。1）垂直方向报文组，是不同会话中偏移为 N 的报文集合 $M_N=\{m_{1N}, m_{2N}, \cdots, m_{xN}\}$，有 N 类报文的协议样本可以生成 N 个垂直方向的报文组；2）水平方向报文组，是来自同一会话中所有偏移的报文集合 $M=\{m_{x1}, m_{x2}, \cdots, m_{xN}\}$。

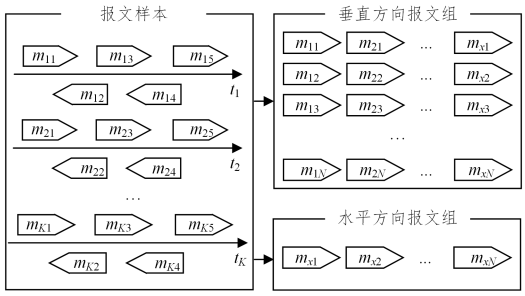

图 9 垂直报文组与水平报文组

Fig. 9 Vertical message groups and horizontal message groups

4.2.4 闭合频繁序列挖掘

挖掘频繁序列是挖掘序列数据库中满足最小支持度阈值的频繁子序列的完全集[19]，效率较低。本文选择挖掘与完全集具有相同效力的闭合频繁序列。常见的序列挖掘算法有BIDE[20]，GSP[21]，SPADE[22]，PrefixSpan[23]。BIDE 算法利用双向扩展方式检测序列闭合性和向后剪枝技术，效率更高。

文献[10]使用协议的首个报文构建样本集，证明了支持度阈值为 0.6 时可以有效提取报文的特征。

输入两方向报文组，为其中的报文字节序列$\langle bt_1, bt_2, \cdots, bt_x \rangle$建立伪投影库，计算它的向前扩展项的个数；循环调用 BIDE 算法，设置支持度阈值为 0.6，计算向后扩展项的个数；若序列$\langle bt_1, bt_2, \cdots, bt_x \rangle$既没有向前扩展项也没有向后扩展项，则其是闭合频繁序列。取所有闭合频繁序列的交集，并将其作为动态域和静态域的划分依据：集合中的序列为静态域，其余字节序列为动态域。

5 密码协议的关键字段语义识别

关键字段的语义识别是协议逆向的重要一步。本文对长度域关键字段和其他关键字段分别采取不同的识别策略。

5.1 长度域字段识别

本文对文献[10]提出的密文长度域识别算法进行优化和改进，利用域划分结果生成长度域取值可能的候选集，在识别密文域的长度字段的基础上对报文长度域、填充字符串长度域等多个长度域进行识别。

5.1.1 生成长度域可能取值候选集

本小节将介绍长度域取值候选集 *LengthValue* 的生成规则。

1)单一域的字节序列长度。将动态域、密文域、明文域的字节序列长度取值加入 *LengthValue* 集合中。

2)静态域与其后动态域或密文域的组合长度。如图 10 所示的报文存在 3 个长度域，指示数据和其关键字段的组合长度。将每个静态域和其后的动态域或密文域组合长度加入 *LengthValue* 集合中。

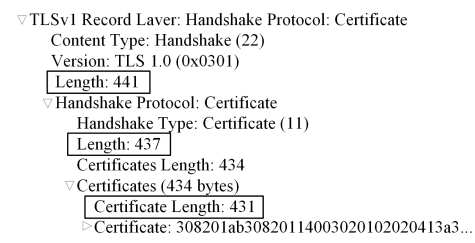

图 10 SSL 协议 Certificate 报文

Fig. 10 Certificate message in SSL protocol

3)动态域和其后静态域的组合长度。密码协议有时在报文末尾使用填充字段来保证报文结构的一致性，如图 11 所示的报文序列中使用 6 个 0x00 字节的填充域。将动态域与其后的静态域的组合长度加入 *LengthValue* 集合中。

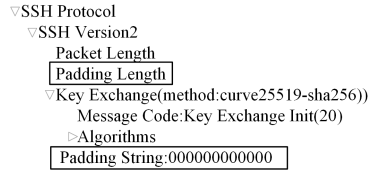

图 11 SSH 协议 KeyExchange Init 报文

Fig. 11 Key Exchange Init message in SSH protocol

4)报文序列长度。报文头部存在一个长度域字段用于指示报文长度，将报文长度加入 *LengthValue* 集合中。

以 ServerHello 报文为例，其生成 *LengthValue* 的具体过程如图 12 所示。

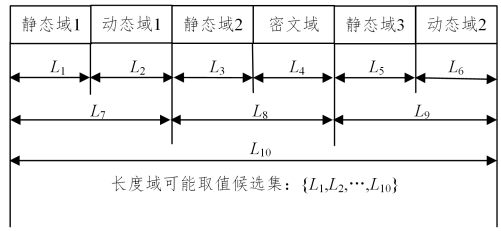

图 12 SSL 协议 ServerHello 报文的 LengthValue 集合

Fig. 12 LengthValue set of ServerHello messages in SSL protocol

5.1.2 循环比对的长度域识别算法

本小节提出的循环比对长度域识别算法针对密码协议中的多个长度域进行识别。基于循环比对的长度域识别算法如算法 1 所示。

算法 1 长度域关键字段的识别算法

输入：(M,(S(a)D(b)C(e)…),LengthValue)

输出：$\{L_1, L_2, \cdots, L_i\}$

1. 初始化 Lkeyword_candidate=$\{\emptyset\}$
2. x_1=oneByteValue(m)
3. x_2=twoBytesValue(m)
4. x_3=threeBytesValue(m)
5. x_4=fourBytesValue(m)
6. for i∈(1,2,3,4) do
7. 找到 ByteValue 相同时 i 最大的字节序列
8. 将x_i加入 Lkeyword_candidate
9. end for
10. for x_i∈Lkeyword_candidate do
11. while x_i满足长度域的启发式规则 do
12. 循环比对(LengthValue,x_i)
13. if LengthValue=x_i then
14. 将x_i加入 Lkeyword
15. end if
16. if LengthValue=x_i-offset then
17. 将x_i加入 Lkeyword
18. end if
19. end while
20. end for
21. return LKeyword

输入报文序列、域划分结果、长度域取值候选集，对序列进行字节片切扫描（见算法 1 第 2－5 行）。依次对静态域和动态域按照 1 字节、2 字节（1 字节和其相邻的前一字节）、3 字节（2 字节和其相邻的前一字节）、4 字节（3 字节和其相邻的前一字节）进行片切并计算其取值分别为x_1，x_2，x_3，x_4。若不同长度片切的字节序列取值相等，则优先选择长度更长的字节序列作为长度域。

进一步判断x_i（$i\in[1,4]$）是否符合长度域的启发规则（见算法 1 第 10 行），若符合则将其加入可能的长度域关键字段集合 *Keywords*。

将集合 *Keyword* 与集合 *LengthValue* 进行循环比对（见算法 1 第 11－16 行），若 *Keyword* 的值或 *Keyword-offset* 与 *LengthValue* 存在交集则判断其为长度域关键字段。

5.2　其他关键字段识别

本小节主要对密码协议特有的关键词字段以及分隔符、序列号和时间戳等关键字段的特点进行总结，如表 3 所列，并依此设计启发策略进行语义识别。

表 3　密码协议关键字段的特点

Table 3　Characteristics of key fields in cryptographic protocols

类别	可打印	定长	变长	静态	动态	特点
分隔符			√	√		取值固定，间隔出现
序列号		√			√	取值不定，在垂直方向报文组递增出现
格式标识		√		√		取值相对固定，位于报文头部
时间戳		√			√	与当前时间相关，长度一般为 4 字节，可以转化为 Unix 时间戳格式
加密算法	√					可打印的 ASCII 字符，包含 SHA，CBC，ASE 等密码学相关字符串
加密套件					√	长度不定，取值可变，内部使用 0x00 分割，数据稀疏

5.2.1　特有关键字段

将密码协议特有的关键字段分为两类，分别设计不同的识别策略。

1）可打印的 ASCII 字符串，通过打印出的字符串来判断关键字段的语义。若字符串包含 SHA，ASE，CBC 等密码学相关字符串，则判断其为加密算法数据字段；若字符串包含 Ubuntu，Windows 等操作系统相关字符串，则判断其为操作系统信息。

2）明文数据，如握手阶段通信双方进行加密方法的协商，发送彼此支持的加密套件。若动态域中字节熵值较大，长度不定，且包含一定数量的“0x00”字节，则判断其为加密套件关键字段。

5.2.2　分隔符或填充字段

分隔符用于划分报文结构，位于多个关键字段之间，取值固定。填充字段用于保证报文格式的一致性。在 EAPOL 协议中，第二次握手需要传输 WPAKeyMIC，其余报文无需此信息，该协议在对应的位置上使用“0x00”进行填充。

统计报文序列中频繁出现的未知固定字节，若不是字符、数字等常用的 ASCII 字节，则判断其为分隔符。查找协议报文中是否有连续出现的分隔符构成的静态域，如有则判定其为填充字段。

5.2.3　序列号

序列号用于标识报文在一个会话中的先后顺序，位于报文头部，取值变化率接近 100%，且与截获报文的先后顺序相对应。

若某动态域在垂直报文组中呈现递增关系，则判断其为序列号字段。

5.2.4　格式标识关键字段

格式标识是用于区别不用类型的报文或者区别报文之后子格式序列的关键字段，如控制码、协议类型、协议版本等。

若位于报头的字节序列在水平方向报文组也不发生改变（不同类型的报文具有相同类型的字节序列），则判断其为协议类型关键字段或协议版本关键字段；若字节序列在水平方向报文组发生变化且取值只有若干可选项（一个取值对应一种子格式序列），则判断其为控制码。

5.2.5　时间戳

时间戳用于测量通信延迟、作为种子生成随机数等，一般使用 UNIX 时间戳（UNIX Time Stamp），长度为 4 字节。

按照 4 字节逐个片切动态域，将其转化为 Unix 时间戳，并将其与提取协议报文的时间戳进行差值计算，若差值结果接近则判断其为时间戳。

6　实验评估

本章对上述方法的有效性进行了实验验证，将逆向分析结果与公开的协议规范进行比较分析，判断域划分的准确性、长度域关键字段提取的完整性和其他关键字段语义识别的正确性，并与相关工作进行比较分析。

6.1　数据集

本文收集了 3 类密码协议的网络流量来构建数据集，分别为安全传输层协议（TLS）、安全外壳协议（SSH）和扩展认证协议（EAPOL）。这些协议的网络流量主要来自于两方面：1）网络上公开的流量数据包；2）使用 Wireshark 和嗅探器抓取实验室局域网环境所产生的网络流量。

对原始数据包进行过滤得到包含 100 条会话的单一类型协议的数据包，使用 Python 和 Scapy 包提取报文序列，构建实验的样本集。样本集的组成具体如表 4 所列。

表 4　样本集构成

Table 4　Composition of dataset

协议	报文类型		报文数量			会话数量
	Client	Server	Client	Server	Total	
SSL	6	2	600	200	800	100
SSH	4	3	400	300	700	100
EAPOL	2	2	200	200	400	100

6.2　结果与分析

得到报文序列后，按字节计算信息熵识别密文域，然后划分动态与和静态域，输出报文的域划分结果。定义 C 为密文域，S 为明文域中的静态域，D 为明文域中的动态域，*Type*

(Length)表示域的长度,长度可变则用 e 表示。接下来以两个报文为例介绍域划分过程及结果分析。

图 13 给出了 SSL 协议的 EncryptedApplicationData 报文的域划分过程,逐字节计算信息熵,确定从第 6 个字节开始为密文域且长度可变,输出 $C(e)$。将密文域替换为字符串"xx",然后将生成报文组输入 BIDE 算法得到水平方向的闭合频繁序列{'00017','00103','00201','005xx'},确定报文的前三个字节为静态域,最终输出的域划分结果为 $S(3)D(2)E(e)$。与报文格式进行比对可发现,密文域变长,密文长度域被划分为动态域 $D(2)$。

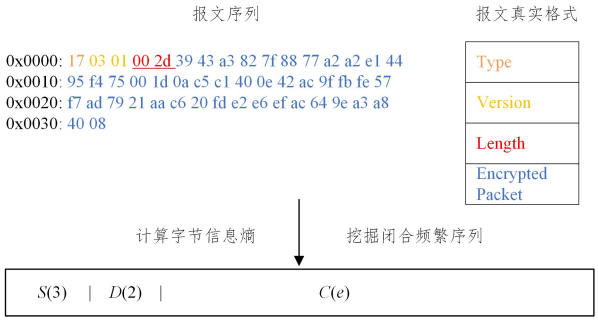

图 13 Encrypted Application Data 报文的域划分过程

Fig. 13 Field partition in Encrypted Application Data messages

图 14 给出了 SSH 协议的 Diffie-Hellman Group ExchangeInit 报文,由信息熵确定其中间的 128 个字节为密文域 $C(128)$。对明文域进行划分,得到域划分结果 $S(10)C(128)S(6)$。由公开规范可知,密文域是定长的共享密钥参数 DH-cliente,其长度域取值固定,被划为静态域。

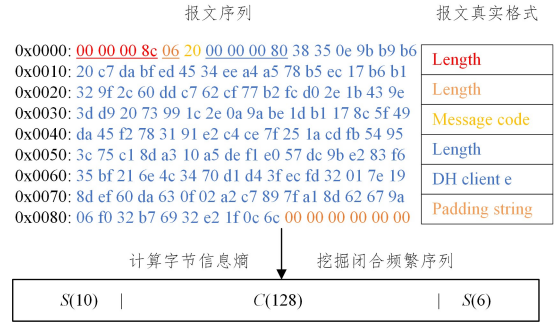

图 14 Diffie-Hellman Group Exchange Init 报文域划分

Fig. 14 Fieldpartition in Diffie-Hellman Group Exchange Init messages

使用本文方法对样本集的其他报文进行域划分,结果如表 5 中的第 3 列所示。SPFPA[10]是一种基于频繁项集的密码协议逆向工具,由于工具未开源,无法直接对本文选取样本进行分析,表 5 中第 4 列是其文献中提到的相同报文的域划分结果。本文方法的域划分结果不仅包括动态域和静态域,同时加入了密文域,并且输出了其长度参数,保证了后续长度域识别的准确性。

但本文方法也存在一些局限,最突出的一个问题是方法会受到样本完备性的影响,同一网络环境下捕获的流量包在某些动态域上取值相同,会对域划分结果产生影响,存在将部分动态域识别为静态域的局限性。

表 5 RAECFS 方法与 SPFPA 方法的域划分结果的对比

Table 5 Comparison between the field identification results of RAECFS and SPFPA

协议	报文	RAECFS	SPFPA
SSH	Protocol	$D(e)$	$D(3)D(e)$
	Key Exchange Init	$S(2)D(3)(1)D(16)S(2)D(e)S(2)D(e)S(2)D(e)$	
	Diffie-Hellman KeyExchange Init	$S(10)C(128)S(6)$	
	New Keys	$S(2)D(2)S(e)$	
	Encrypted packet	$C(e)$	
SSL	Client Hello	$S(3)D(2)S(1)D(3)S(2)D(4)C(28)D(1)C(e)S(1)D(e)$	$S(3)D(2)S(1)D(e)$
	Certificate	$S(3)D(2)S(1)D(6)C(e)$	
	Client Key Exchange	$S(3)D(2)S(2)D(2)C(e)$	
	Change Cipher Spec	$S(5)D(1)$	
	Encrypted HandshakeMessage	$S(3)D(2)C(e)$	
	Application Data	$S(3)D(2)C(e)$	
EAPOL	1/4	$S(17)C(32)S(50)$	
	2/4	$S(17)C(32)S(32)C(16)D(e)$	
	3/4	$S(17)C(32)S(32)C(16)D(2)C(e)$	
	4/4	$S(2)D(2)S(76)\ C(16)S(2)$	

6.2.1 长度域关键字段识别

本文使用识别率来评价该长度域识别算法的识别效果。定义样本集中所有被正确识别的长度域关键字段数量为 N,样本集中长度域关键字段总数为 C,识别率为 N/C,识别率越高说明算法的识别效果越好。结果如表 6 所列,该算法对 SSL 协议和 SSH 协议中的长度域关键字段识别率超过 80%,而 EAPOL 协议由于采用'0x00'对未使用的关键字段进行占位保留,识别率较低。

表 6 长度域关键字段识别率

Table 6 Identification rates of length fields

协议	C	N	识别率 N/C/%
SSL	19	16	84.21
SSH	25	22	88.00
EAPOL	12	7	58.33

接下来以两种报文为例说明长度域关键字段识别过程及结果分析。

SSL 协议 Encrypted Application Data 报文的长度域关键

字段识别过程如图 15 所示，根据域划分结果生成长度域取值候选集$\{e,e+2,e+5\}$，按 1～4 字节对动态域和静态域进行片切，循环比对字节取值，发现 $byteValue(D(2))=e$，确定第 4 和第 5 字节一起构成了长度域关键字段，指示其后密文域 $C(e)$的长度。

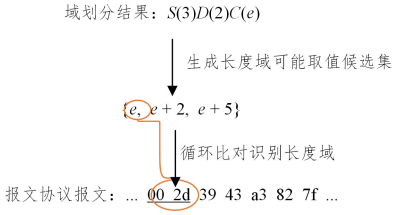

图 15　Encrypted Application Data 报文长度域识别过程

Fig. 15　Length field identification for Encrypted Application Data message

EAPOL 协议的第二次握手报文的长度域识别过程如图 16 所示。

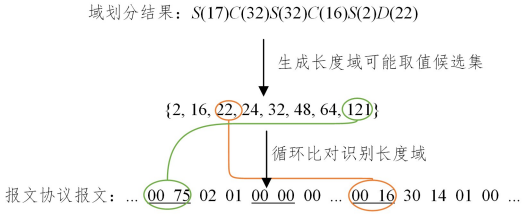

图 16　EAPOL 协议报文长度域识别过程

Fig. 16　Length field identification for EAPOL

根据域划分结果，生成长度域取值候选集{32，16，2，22，24，48，64，121}，循环比对后发现 $bytevalue(S(2))=22$，$bytevalue(offset(4))=121-offset$，确定偏移为 3～4 字节与静态域 S(2)为长度域，但由于样本的不完备性，偏移 7～8 的长度域“0000”被划分为静态域，无法被准确识别。

由于该算法对字段的片切边界与域划分边界相同，因此当某长度域字段被切分为静态域和动态域时，无法进行有效识别。将本文方法与 SPFPA 方法对相同报文的解析结果进行对比，结果如表 7 所列，本文方法可以对长度域字段进行更加全面、准确的识别。

表 7　RAECFS 与 SPFPA 方法的长度域识别结果对比

Table 7　Comparison between the length field identification results of RAECFS and SPFPA

协议报文	关键字段	RAECFS	SPFPA
SSL Client Hello	Protocol Type	√	√
	Protocol Version	√	√
	Length	√	×
	Handshake Type	√	√
	Length	√	×
	Protocol Version	√	√
	Random	×	×

6.2.2　关键字段的语义识别

下面给出协议关键字段语义识别的实验结果，并将结果与 Wireshark 内的解析结果或者公开的协议规范进行对比。

对 EAPOL，SSL，SSH 协议的部分报文关键字段语义的识别结果如图 17 所示。EAPOL 协议的第二次握手报文，利用水平方向报文组的闭合频繁序列可以确定一些控制字段，如 Version 和 Type，但无法识别比特级别的关键字段语义，如 Keyinformation 等。SSH 协议的 Encryptedpacket 报文存在长度域、校验码和加密数据包信息，但全部是密文域，因此无法对其进行有效的域划分和语义识别。

EAPOL-2/4

Version	√
Type	√
Length	×
Type	√
Key information	×
Length	×
Replay Counter	×
WPA Key Nonce	√
Key IV	×
WPA Key RSC	×
WPA Key ID	×
WPA Key MIC	×
Length	√
WPA Key Data	√

SSL-Certificate

Content Type	√
Version	√
Length	√
Handshake Type	√
Length	√
Certificates Length	√
Certificate Length	√
Certifate	√

SSH-Diffe Hellman Key Exchange Init

Length	√
Length	√
Message Code	√
Length	√
DH Client e	√
Padding String	√

SSH-Encrypted packet

Packet Length	×
Encrypted Packet	×
MAC	×

图 17　部分协议报文的语义识别结果

Fig. 17　Results of semantic recognition of messages

结束语　本文提出的密码协议逆向分析方法首先通过信息熵和闭合频繁序列对报文进行域划分，然后利用生成取值候选集的循环比对算法实现对密码协议多种类型长度域的识别，使用启发策略对密码协议一些特有关键字段进行语义识别。最后使用 3 类密码协议对方法的有效性进行了验证，结果表明了本文方法的有效性。进一步的研究工作包括样本不完备情况下的逆向分析准确率提升方法、密码协议状态机推断等。

参 考 文 献

[1]　WANG Z F，CHENG G，MA W J，et al. Research progress of network protocol reverse engineering technologies based on network trace [J]. Journal of Software，2022，33(1)：254-273.

[2]　KLEBER S，MAILE L，KARGL F. Survey of protocol reverse engineering algorithms：decomposition of tools for static traffic analysis[J/OL]. IEEE Communications Surveys & Tutorials，

2018. https://ieeexplore.ieee.org/document/8449079.
[3] WU L F, HONG Z, PAN F. Network protocol reverse analysis and application[M]. Beijing: National Defense Industry Press, 2016.
[4] YE Y, ZHANG Z, WANG F, et al. Netplier: probabilistic network protocol reverse engineering from message Traces[C]// Network and Distributed System Security Symposium. 2021.
[5] GENTRY C, WATERS B. Adaptive security in broadcast encryption systems(with short ciphertexts) [C]// Annual International Conference on the Theory and Applications of Cryptographic Techniques. 2009: 171-188.
[6] ZHAO X, ZHANG F. Fully CCA2 secure identity-based broadcast encryption with black-box accountable authority[J]. Journal of Systems and Software, 2012, 85(3): 708-716.
[7] SHI X L, ZHU Y F, LIU L, et al. Method of encrypted protocol reverse engineering[J]. Application Research of Computers, 2015, 32(1): 214-217.
[8] GAO J F, ZHANG Y F, LUO S, et al. Research on Taint Backtracking Reverse Analysis Method of Network Encoding Protocol[J]. Netinfo Security, 2017(1): 68-76.
[9] MA R K, ZHENG H, WANG J Y, et al. Automatic protocol reverse engineering for industrial control systems with dynamic taint analysis[J]. Frontiers of Information Technology & Electronic Engineering, 2022, 23(3): 351-360.
[10] ZHU Y, HAN J, YUAN L, et al. SPFPA: A format parsing approach for unknown security protocols[J]. Journal of Computer Research and Development, 2015, 52(10): 2200.
[11] HE X D. Security Analysis of Security Protocol Implementations Based on Network Trace [D]. Wuhan: South-Central Minzu University, 2019.
[12] TANG S Y, CHENG G, JIANG B M, et al. Detection and recognition of VPN encrypted traffic based on segmented entropy distribution[J]. Cyberspace Security. 2020, 11(8): 23-27, 33.
[13] XIAO D Q, ZHOU Q, ZHANG H G, et al. Analyzing encryption protocols based on temporal logic[J]. Chinese Journal of Computers, 2002, 25(10): 1083-1089.
[14] DING S F, ZHU H, XU X Z, et al. Entropy-based fuzzy information measures[J]. Chinese Journal of Computers, 2012, 35(4): 796-801.
[15] ZHU Y N, HAN J H, YUAN L, et al. Protocol ciphertext field identification by entropy estimating[J]. Journal of Electronics & Information Technology, 2016, 38(8): 1865-1871.
[16] FELFMANN A, ZITTERBART M, CROWCROFT J, et al. Technologies, Architectures, and Protocols for Computer Communication[C]// ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. 2003.
[17] OLIVIAN J, GUOBAULT-LARRECG J. Detecting subverted cryptographic protocols by entropy checking[D]. LSV, ENS Cachan, 2006.
[18] KLEBER S, MAILE L, KARGL F. Survey of protocol reverse engineering algorithms: Decomposition of tools for static traffic analysis[J]. IEEE Communications Surveys & Tutorials, 2018, 21(1): 526-561.
[19] WANG H, DING S F. Research and development of sequential pattern mining(SPM)[J]. Computer Science, 2009, 36(12): 14-17.
[20] WANG J, HAN J. BIDE: efficient mining of frequent closed sequences[C]// Proceedings 20th International Conference on Data Engineering. 2004: 79-90.
[21] SRIKANT R, AGRAWAL R. Mining sequential patterns: Generalizations and performance improvements[C]// International Conference on Extending Database Technology. 1996: 1-17.
[22] ZAKI M J. SPADE: An efficient algorithm for mining frequent sequences[J]. Machine Learning, 2001, 42(1): 31-60.
[23] PEI J, HAN J, MORTAZAVI-ASL B, et al. Mining sequential patterns by pattern-growth: The prefixspan approach[J]. IEEE Transactions on knowledge and data engineering, 2004, 16(11): 1424-1440.

LIANG Chen, born in 1998, postgraduate. Her main research interests include cybersecurity and reverse engineering.

WU Lifa, born in 1968, Ph. D, professor, Ph.D supervisor. His main research interests include cybersecurity and software security.

(责任编辑:喻藜)