未知二进制协议的报文分割方法

徐　魁[1],海　洋[1],李晓辉[1],朱承才[2],陶　军[2,3]
(1. 宝鸡市公安局通信处,陕西 宝鸡 721014;
2. 东南大学 网络空间安全学院,江苏 南京 211189;
3. 计算机网络和信息集成教育部重点实验室(东南大学),江苏 南京 211189)

摘　要:基于网络轨迹的协议逆向工程使用捕获的数据包进行分析,进而逆向未知协议的格式等信息。该文提出了一种利用二进制协议在网络通信过程中使用报文序列数据集来推断消息字段划分的新方法 HV。该方法首先利用定义的测度分析各条消息中的值分布,分析报文的内部结构,对字段边界初次划分。接着利用消息序列之间所隐藏的统计信息对字段边界再次划分。最后将两次划分的结果结合,生成最终的字段划分结果。此前的研究很少利用每个消息内部的结构特征,而是通过比较多条消息得出结论。对于消息之间的统计特征,该文仅仅比较相邻的消息,而不是相互比较多条消息。此外,该文还定义了格式匹配分数,用于消息字段划分的质量的度量。将格式匹配分数应用于 HV 和以前的方法的对比实验中,进而验证 HV 字段划分的质量。由于 HV 在水平分析上利用了消息的内部结构,并且在垂直分析中只比较相邻消息之间的异同,因此 HV 不仅具有较好的字段划分效果,而且只有线性复杂度。

关键词:二进制协议;协议逆向;字段划分;报文格式;内在结构
中图分类号:TP393　　**文献标识码**:A　　**文章编号**:1673-629X(2023)11-0119-07
doi:10.3969/j.issn.1673-629X.2023.11.018

Message Segmentation Method of Unknown Binary Protocol

XU Kui[1],HAI Yang[1],LI Xiao-hui[1],ZHU Cheng-cai[2],TAO Jun[2,3]
(1. Communication Office of Baoji Public Security Bureau,Baoji 721014,China;
2. School of Cyber Science and Engineering,Southeast University,Nanjing 211189,China;
3. The Key Laboratory of Computer Networks and Information Integration,Ministry of Education,
Southeast University,Nanjing 211189,China)

Abstract:Protocol reverse engineering based on network trajectory uses captured data packets to analyze and reverse information such as the format of unknown protocols. A new method,HV,is proposed to infer the division of message fields by using the message sequence data set used by binary protocol in network communication. HV uses the defined measure to analyze the value distribution in each message and then analyze the internal structure of the message,so that the field boundary can be divided for the first time. Then HV uses the hidden statistical information between message sequences to divide the field boundaries again. Finally,the results of the two divisions are combined to generate the final field division result. Previous studies rarely use the internal structural features of each message,but draw conclusions by comparing multiple messages. For the statistical characteristics between messages,we only compare adjacent messages,rather than comparing multiple messages with each other. In addition,we also define the format matching score,which is used to measure the quality of message field division. Applying the format matching score to the comparison experiment between HV and previous methods,the quality of HV field division is verified. Because HV uses the internal structure of messages in horizontal analysis and only compares the similarities and differences between adjacent messages in vertical analysis,HV not only has good field division effect,but also has linear complexity.

Key words:binary protocol;protocol reverse;field division;message format;internal structure

0　引　言

随着互联网技术的不断发展,保障通信网络的安全愈发重要。二进制协议因其非侵入性的特点以及在网络中的广泛应用成为了当前研究的热点。

收稿日期:2023-01-04　　**修回日期**:2023-05-04
基金项目:中国高校产学研创新基金-阿里云高校数字化创新专项(2021ALA03006)
作者简介:徐　魁(1973-),男,高级工程师,研究方向为大数据分析、信息检索;通信作者:陶　军(1975-),男,博士,教授,博导,CCF 高级会员(42637S),研究方向为网络安全、物联网技术等。

协议逆向工程的关键在于对协议报文如何分段。对未知规范的通信需要通过监控网络流量等方法对其进行逆向工程[1]。静态流量逆向工程的应用包括僵尸网络分析[2]、蜜罐设置[2-3]、模糊漏洞测试[4]和网络自动建模[5]。2004 年,Beddoe[6]和 Rauch 提出了首个解决方案:序列比对。此后 ProDecoder[7]和 PRISMA[2]发现自然语言处理在使用 ASCII 编码关键字来构造消息的协议上运行良好。推断消息格式往往需要大量消息,然而多序列比对会导致指数复杂性[8],当数据量巨大时性能变差。另一方面,长可变消息可能出现对齐偏差,从而导致字段边界误判。因此,为了聚类消息类型或对齐字段序列,就需要进行消息字段划分。ScriptGen[3]、Discoverer[9]和 Netzob 使用序列对齐来推断消息格式。FieldHunter[10-11]使用字段类型的特征化等方法进行格式推断。张蔚瑶等人使用协议特征库对未知协议进行逆向分析[12]。此外,研究人员提出了三类创新性的协议报文分段方法:基于信息论投票的报文分段、基于决策模型的报文分段[13]与基于报文内部结构的报文分段。Zhang 等人[14]提出协议关键词提取方法 ProWord,首次将无监督专家投票算法应用于流量分析。Sun 等人[15]引入统计信息,从信息论的角度提出协议报文分段算法 ProSeg。IPART[16]在专家投票算法基础上又加入语义识别,对报文分段点进行二次确认。Jiang 等人[17]提出基于相邻字节距离的报文分段算法 ABInfer,采用最近邻聚类算法迭代将相邻字节进行合并,然后对字段进行划分。

协议字段划分的过程可以抽象为报文字节序列中字段边界的决策问题。黎敏等人[18]将字段划分过程看成马尔可夫过程,在此基础上使用隐半马尔可夫模型(Hidden Semi-Markov Models,HSMM)[19]进行字段划分。Cai 等人[20]同样使用隐半马尔可夫模型进行求解,对黎敏的工作进行了优化。Tao 等人[21]使用贝叶斯决策模型进行协议逆向分析,提出了对二进制协议进行字段划分的方法 PRE-Bin。

协议报文的部分研究以比特为粒度,挖掘比特间的表征关系。Kleber 等人[22]研究协议报文的内部结构,提出了一种新颖的报文分段方法 NEMESYS。Marchetti 等人[23]通过幅度序列和位翻转频率寻找报文分段点,提出汽车通信数据帧分段方法 READ。

基于上述情况,该文提出了一种用于未知二进制协议逆向工程的协议字段划分方案 HV。主要工作如下所述:

首先,提出字节翻转率的概念并将其应用到消息分析。从垂直分析的角度,通过对比相邻消息的结构,找到该二进制协议在消息结构上的共性。其次,从水平分析的角度探究单条消息的内部结构。基于第一数字定律等方法初步找到消息边界;使用路径搜索等算法找到更多候选边界点,从而优化消息字段划分的结果。接着,创新性地联合水平以及垂直分析进行消息字段的划分,设计用于未知二进制协议字段划分方案 HV。对从上述得到的消息分段点进行评估、投票等决策,得到最终结果。最后,引入格式匹配分数(Format Match Score,FMS)用于量化特定消息的格式推断质量。

1 算法思路

1.1 基于垂直分析的报文分段

此阶段探究的是消息之间所呈现出的结构信息。对相邻的消息进行比较,得出相关的统计信息。

1.1.1 字节翻转率与位翻转率

一般工业协议粒度为字节,将消息载荷以字节形式展开,使用字节翻转率进行评估。字节翻转率定义如下:

$$BF_i = \frac{\sum_{m_j \in M} m_{j+1}(i) \oplus m_j(i)}{|M| - 1} \tag{1}$$

其中,BF_i 表示第 i 个字节的翻转率,M 是所有消息集合,m_j 是 M 集合中第 j 条消息,$m_j(i)$ 是第 j 条消息的第 i 个字节。$\oplus$ 是异或操作。$|M|$是消息集合中的消息数量。这一步得到一个含有 n 个元素的数组,每个元素代表某一字节处的翻转率,n 是消息载荷字节的长度。字节翻转率独立于同一消息中邻近的字节,只与邻近的消息有关。过程如算法 1 所描述。

算法 1:字节翻转率计算

```
Function calculateBitFlip(Messages,payloadLen):
    messagesNum←len(Messages);
    byteFlip array(payloadLen);
    FOR byteIndex IN range(0 , payloadLen):
        FOR mIndex IN range(1 , messageNum):
            IF message[mIndex][byteIndex]
                    ≠ message[mIndex][byteIndex-1]:
                byteFlip[byteIndex]+=1;
    return byteFlip;
```

同理,可以将消息载荷以比特形式展开,得到位翻转率。定义如下:

$$bF_i = \frac{\sum_{m_j \in M} \mathrm{bit}(m_{j+1})(i) \oplus \mathrm{bit}(m_j)(i)}{|M| - 1} \tag{2}$$

位翻转率处理的粒度是比特位。

1.1.2 字段划分

对消息进行翻转率的计算后可以得到字节翻转率数组 BF 以及位翻转率数组 bF。接着进行字段划分。

首先遍历字节翻转率数组,查找符合如下条件之一的字节位置:

(1)该位置字节的翻转率为局部极值点,即满足:$BF_i \geqslant BF_{i-1}$ and $BF_i \geqslant BF_{i+1}$。

(2)该位置字节与相邻的位置字节都具有一个较高的翻转率,即满足:

$BF_i \geqslant \Phi$ and ($BF_{i-1} \geqslant \Phi$ or $BF_{i+1} \geqslant \Phi$)。

(3)该位置字节的翻转率为 0,即 $BF_i = 0$。

将符合条件的字节位置标记为字段边界可疑点。经过上述处理得到一个边界列表 b。将字节翻转率为 0 的字节位标记为边界。根据翻转率的定义,经常变化的字段翻转率会偏大,反之则偏小。

位翻转率数组是一个辅助数组。对于计数字段,翻转率会较大。计数字段低位的字节翻转率为 1,相应的最低比特位翻转率也为 1。并且计数字段从最低位到最高位翻转率应是递减的,每一位的翻转率是下一位的两倍。当字节翻转率为 1 后可利用位翻转率数组进行确认。当认定某一字节为计数字段时,需要查看该字节的后一字节以及前一字节。这一过程如算法 2 所描述。

算法 2:字段划分阶段二

```
function divide2(BF, bF, b):
    BLen len(BF);
    bLen len(bF);
    For i In range(0, BLen):
        IF BF_i = 1:
            IF detectLeftMatch(bF, i):
                b.append((i-2, BF_{i-2}));
                b.erase(i-1);
            ELIF detectRightMatch(bF, i):
                b.append((i+1, BF_{i+1}));
                b.erase(i);
    return b;
```

通过算法 2 可以得到垂直分析的边界列表 b。该阶段是在寻找同一种协议的所有消息中共有部分的统计特性。在进行消息分段时需要取一个固定的长度进行消息间比对。具体如何取值下文有所说明。

1.2　基于水平分析的报文分段

协议字段划分的过程可以抽象为报文字节序列中字段边界的决策问题。因此可以使用路径搜索算法从水平分析的角度对消息的内部结构进行分段。在此之前需要进行分支度量以及约束条件的定义。

1.2.1　分支度量的定义及第一数字定律扩展

第一数字定律,指所有自然随机变量只要样本空间足够大,每一样本首位数字为 1 至 9,各数字的概率在一定范围内具有稳定性。以 1 为首位数字的数的出现概率约为总数的三成。总结而言,越大的数以它为首几位的数出现的概率就越低。

在十进制中,以 n 开头的数出现的几率为:

$$P(n) = \log_{10}(1 + \frac{1}{n}), n \in [1,9], n \in N^+ \quad (3)$$

然而二进制协议中对以“0”开头的字节也会保留,因此可以扩展为:

$$P(n) = \log_{10}(1 + \frac{1}{n+1}), n \in [0,15], n \in N \quad (4)$$

其中,n 所有取值的概率和为 1,即:$\sum_{n=0}^{15} P(n) = 1$。

分支度量在定义时主要基于第一数字定律,边界评估指标如式(5)所示:

$$\text{score}_i = B_{i-1}^{1,2} \cdot P(B_i^1) \cdot \alpha \quad (5)$$

其中,score_i 表示第 i 个候选边界的评估分数,P 源于表达式(4),$B_i^{1,2}$ 表示消息载荷中的第 i 个字节,右上角数字代表字节的前 4 比特或者后 4 比特,分别用 1 和 2 标识。例如某条消息第 2 个字节为 0f,则可以表示为 $B_2^{1,2}$,则 B_2^1 为 0,B_2^2 为 f。α 是标准化系数,取值为 0.016。为了体现前后字节的落差,该式子乘以 $B_{i-1}^{1,2}$ 进行乘法扩大。

1.2.2　约束条件

约束条件控制节点之间是否可达。构造约束条件时:首先字段长度是有限的,一般不超过 4 个字节,个别字段会达到 8 字节,多数情况下为偶数或“1”。其次,字段是单向的,即字段是从左往右,不存在从右往左的。评估指标如式(6)(7)所示:

$$d_{i,j} = \frac{1}{\text{score}_j} \cdot w_{j-i} \quad (6)$$

$$w_k = \begin{cases} -\infty (k \leqslant 0 \text{ or } k = 3 \text{ or } k > 4, k \in Z) \\ \dfrac{k^2}{1 + 2^2 + 4^2} (k = 1,2,4) \end{cases} \quad (7)$$

式(6)中,$d_{i,j}$ 表示第 i 个候选边界和第 j 个候选边界之间的距离。根据最短路径搜索的思想,为使最佳路径的路径权值和最小,该式使用 score 的倒数作为分支度量。式(7)中,w_k 是一个距离权重,其中 k 是整数,表示两个候选边界之间相隔的字节数。当 $k = 1,2,4$ 时,表示 $j > i$ 并且字段长度合理,使用平方增量;当 $k \leqslant 0$(从右往左)或者 $k = 3$(3 个字节长度的字段一般不常见)或者 $k > 4$(字段长度太长)时,权重为负无穷,即不可达。

1.2.3　最佳路径搜索算法

根据分支度量和约束条件生成候选边界有向图,利用最佳路径搜索算法从有向图中找到与真实格式关

键词边界最接近的一条路径作为最终格式关键词的边界推断结果。目标函数如式(8)所示：

$$\arg\min_{\text{trace}_k}\{\sum d_{i,j}\}, d_{i,j} \in \text{trace}_k, \text{trace}_k \in \text{Trace} \quad (8)$$

其中，Trace 是所有可能的路径集合，trace_k 是集合的第 k 条路径。

最佳路径搜索算法的关键是：根据 $d_{i,j}$ 推测当前路径上最后一个边界点 b_i 的可能的下一个候选边界 b_j。由最佳路径 $B_{\max}=\{b_{I0},b_{I1},\cdots,b_{Ik}\}$ 寻找到达下一个边界 b_j 的可行路径，并计算各可行路径的分支度量总和 $\sum_{i=1}^{k} d_{I_{i-1},I_i}$，直到完成对 b_j 所有可行路径的遍历。最后选择分支度量总和最小的一条路径作为到节点 b_j 的最佳路径。这里采用 01 编码存储边界情况，0 代表不是边界，1 代表是边界。报文开头和结尾默认是边界。

最佳路径搜索算法应用于关键词边界选择，最终目标是寻找一条从第一个候选边界点到最后一个候选边界点权值之和最小或最大的路径。

1.3 联合垂直分析和水平分析的消息分段

1.3.1 划分方案

消息的水平分析通过消息内部所蕴含的信息对字段进行了划分。消息的垂直分析通过消息序列之间所蕴含的信息对字段进行了划分。这两种方案各自的特点如表 1 所示。

表 1 水平与垂直分析优缺点

	特点	优点	缺点
水平分析	分析对象为单条消息。分析所使用的数据集不必是有序的。利用消息内部字节之间所蕴含的信息	可以利用消息内部信息。对于消息数量要求较低。算法复杂度低	对每条消息单独分析，获取的信息有限，很局限
垂直分析	分析对象为整个消息数据集。分析使用的数据集有序。利用不同消息之间相同位置处的字节或者比特信息	可以利用多条消息的统计信息，得出的结果更具普遍性。能得出一些更深层次的字段信息，如常量，计数字段	对于消息数量以及质量要求高，数据量少时效果欠佳

该文将这两种方式进行结合，设计了一种创新性的消息字段划分方案 HV。

对于水平分析，分析的是单条消息；结合垂直分析时要将分析对象由单条消息转为消息集合。

首先，消息字段的不等长导致有的消息会推导出边界，有的则没有，因此主要比较共有部分。此时需要在水平分析的基础上加入投票机制。图 1 对投票机制进行了演示。实例中有三条消息，消息的 data 字段补充 xx 来对齐消息，每条消息都有相应的字段划分边界。如图 1 所示，每一个候选分段点都有相应的票数；然后遍历投票结果 vote，对每一个候选字段划分边界的得票率 $\frac{\text{vote}_i}{|M|}(0 \leqslant i < \text{len}(\text{vote}))$ 与一个事先设定好的阈值 Θ 进行比较，当得票率低于阈值时就是一个伪字段边界划分点；否则就推断为字段边界划分点。

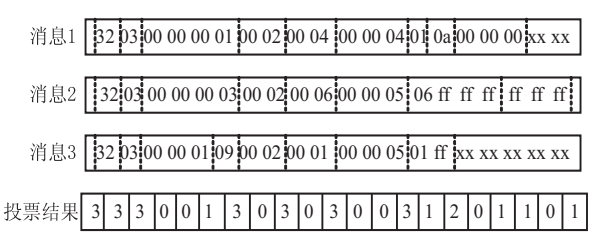

图 1 投票机制执行过程

接下来，结合垂直分析与水平分析的结果。设水平分析的字段边界划分结果为 $I_h=\{i_{h,1},i_{h,2},\cdots\}$，垂直分析的字段边界划分结果为 $I_v=\{i_{v,1},i_{v,2},\cdots\}$，最终的结果取两者的并集，即 $I=I_h \cup I_v$。

1.3.2 方案优化

为避免在 data 字段进行消息字段划分对结果产生干扰，需要对消息进行截尾处理。这里使用平均类内距离作为评估指标，如式(9)所示：

$$\text{avgDis}=\frac{|M|^2-|M|}{2\text{len}(m_i)}\cdot\sum_{m_i\neq m_j\in M} d(m_i,m_j) \quad (9)$$

其中，$d(m_i,m_j)$ 是两个消息字段划分后的 01 向量之间的曼哈顿距离，$\frac{|M|^2-|M|}{2\text{len}(m_i)}$ 是归一化因子。

接着对消息作截尾处理，取不同长度的消息计算平均类内距离，取平均类内距离骤增时的消息长度 len_{avg} 作为截尾点候选点。同时考虑所有消息中最短的消息长度 $\text{len}_{\min}$。将 len_{avg} 的初始值设为所有消息的长度最小值，再设一个下限 len_{low}，在该文中设为 10。最终的截尾长度 $\text{len}_{\text{final}}$ 取值如下：

$$\text{len}_{\text{final}}=\min\{\text{len}_{\min},\max\{\text{len}_{\text{avg}},\text{len}_{\text{low}}\}\} \quad (10)$$

1.4 格式匹配分数 FMS

引入格式匹配分数 FMS 作为字段划分质量的度量。该测度主要考虑三个方面：(1)正确识别字段的比率；(2)区分移位字段边界和完全错误字段；(3)量化不同字段边界推断的递减效用。

FMS 为消息的每一个真实边界 r_k 定义了范围，一个边界的范围起始点为前一个边界 r_{k-1} 和前一边界 r_k 的中间点，范围结束点为当前边界 r_k 和后一边界 r_{k+1} 的中间点。消息开始处 r_0 和消息结束处 $r_{|R|}$ 不分配边界范围。因此，当推断边界 i_l 满足式(11)时，就表

明 i_l 属于 r_k 的范围。

$$r_{k-1}+\frac{r_k-r_{k-1}}{2}\leqslant r_l\leqslant r_k+\frac{r_{k+1}-r_k}{2} \tag{11}$$

式中,i_l 表示第 l 个推断的字段边界的下标索引,$0<l<|I|$,其中 $|I|$ 是推断出的边界数。r_k 是第 k 个真实边界的下标索引,$0<k<|R|$,其中 $|R|$ 是真实边界数。定义 δ_r 为真实边界 r_k 到最近的推断的边界 i_l 的距离,如式(12)。其中 i 满足式(11)。

$$\delta_r=\operatorname{argmin}\{|i-r|\}-r \tag{12}$$

将空集上的 min 运算符定义为 $\min\phi=-\infty$。可知 δ_r 有四种情况:① $\delta_r=-\infty$:对于真实边界 r,没有与之匹配的推断结果。② $\delta_r=0$:推断边界与真实边界完全吻合。③ $-\infty<\delta_r<0$:推断边界在真实边界的左边,偏移量为 δ_r 字节。④ $\delta_r>0$:推断边界在真实边界的右边,偏移量为 δ_r 字节。

模式匹配分数的定义如式(13)所示。

$$\mathrm{FMS}=\frac{1}{|R|}\sum_{r\in R}e^{\left(-\left(\frac{\delta_r}{\gamma}\right)^2\right)}\cdot e^{\left(-\frac{||SR_r|-|SI_r||}{|SR_r|+|SI_r|}\right)} \tag{13}$$

FMS 中使用高斯权重对每个边界匹配进行非线性加权。给距离 δ_r 分配一个测度 $e^{\left(-\left(\frac{\delta_r}{\gamma}\right)^2\right)}$,精确匹配时权重为 1,缺少相应的边界时权重为 0。通过调整参数 γ 可以调整 FMS 随推断边界与真实边界位置的偏移量增加时下降的陡度。

FMS 中 $e^{\left(-\frac{||SR_r|-|SI_r||}{|SR_r|+|SI_r|}\right)}$ 是偏移量惩罚,推断出的分段数与真实分段数的偏差越大,这个量就越小。

最后,对整个式子进行标准化,使得 FMS 的值在 0 到 1 之间。推断质量越高,FMS 越大。

2　实　验

2.1　实验设置

实验样本为 668 条 S7COMM 协议数据、115 条 DNP3.0 协议数据、3 948 条 Modbus 协议数据和 674 条 EGD 协议数据。S7COMM 协议使用 TPKT 和 COPT 封装 PDU,DNP3.0 协议较复杂,Modbus 协议较简单,EGD 协议含有时间戳等字段。

由于实验无需对不同协议的字段划分结果进行横向比较,因此并未保持协议数据的数据量一致。实验中投票时的 Θ 设置为 0.8,FMS 的 γ 设置为 2,使用 tshark 对消息解析出的信息作为基准。

2.2　实验结果及性能分析

图 2 展示了四种协议截尾后的投票结果。可见划分结果的有效字段较多,侧面印证截尾的必要性。

该阶段对字段边界的推导初具成效,尤其是 EGD 协议。这是因为 EGD 协议主要由 IP 地址和时间戳字段等常见字段组成。这些字段在投票前已经被事先定义的字段识别方法识别了。

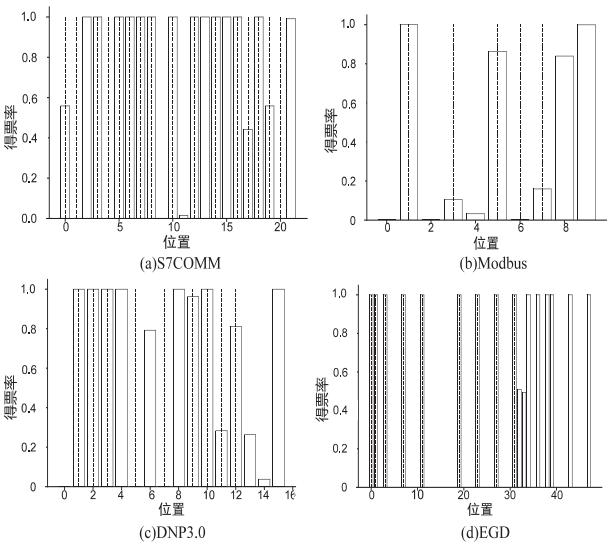

图 2　投票结果

在垂直分析时需要计算位翻转率,图 3 展示了这四种协议的位翻转率。可见字段边界左边的比特位的翻转率普遍偏高,所以翻转率高的位置附近或者翻转率骤降点可以判定为字段边界。

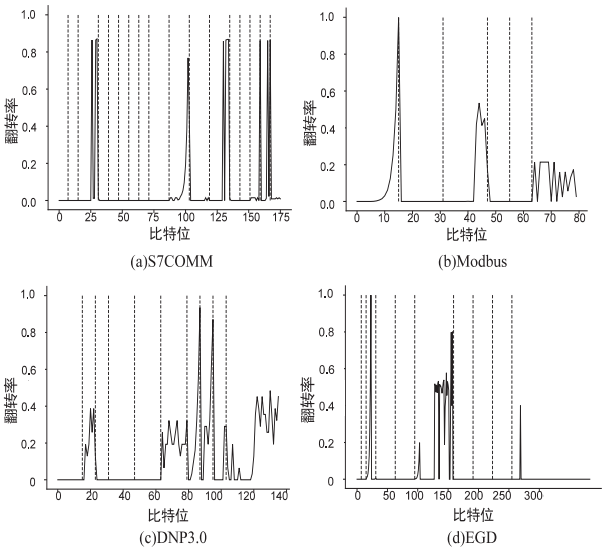

图 3　位翻转率

实验使用 Netzob 以及 NEMESYS 作为对比方法,如图 4 所示。(a)描述的是 S7COMM 协议的字段划分的质量,可见随着数据量的增加三者的质量变化都较小。(b)描述的是 Modbus 字段划分的质量,其中 Netzob 推测的质量最高。(c)描述的是 DNP3.0 字段划分的质量。可见 Netzob 的质量较差。因为实验中 DNP3.0 协议的数据量太少,Netzob 可挖掘的统计信息太少。且 NEMESYS 和 HV 都有较大起伏,这是因为 NEMESYS 只考虑单条消息,依赖于每条消息的取值,所以不受样本数增加的影响;但是 HV 在考虑单条消息的同时也会考虑多条消息之间的比较,因此随着相似样本的增加会提高推测质量。(d)描述的是 EGD

字段的划分质量,清晰地观察到 HV 的推导质量极高且稳定,达到了 0.725,这是因为 EGD 协议中有几个字段的语义被 HV 事先定义了。

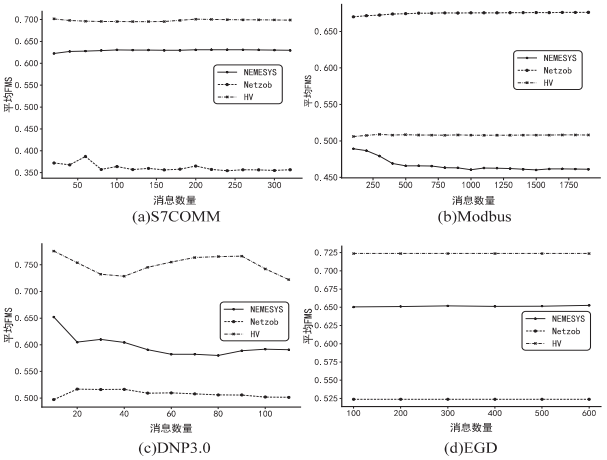

图 4 FMS 测度下的字段划分质量

图 5 展示了 HV 与 Netzob 和 NEMESYS 在所有消息上的推导质量的分布情况。从图中可以看出 HV 除了在 Modbus 协议上的推测质量不如 Netzob,其余情况下的推测质量均高于 Netzob 和 NEMESYS,而且较为稳定。HV 整体上是优于 Netzob 以及 NEMESYS 的。

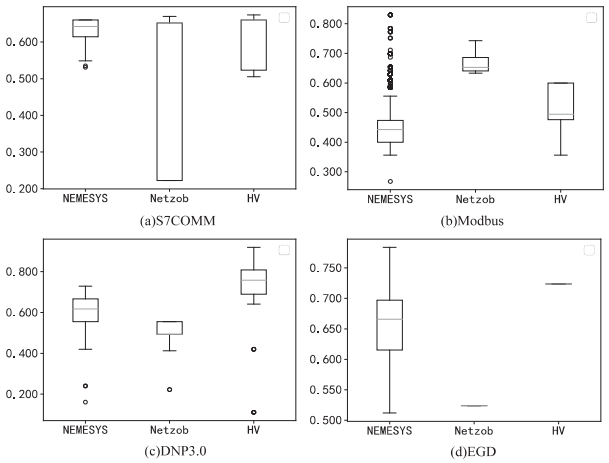

图 5 FMS 分布情况

表 2 列出了三种方法的运行时间。其中+∞表示在 30 分钟内无法求解。可以看出,三种方法中 Netzob 的运行时间最长,甚至当数据集到达一定数量时,可能无法求解。这是由于 Netzob 和大多数协议逆向算法相同,使用了全局序列比对算法,导致具有指数级别的时间复杂度。当对最大长度为 l 的 k 条消息进行比对时,复杂度为 $O(l^k)$。从中还可以看出,NEMESYS 运行时间最短,这是因为 NEMESYS 不需要将数据集中的消息进行任何比较,它只与消息的长度以及数量相关,导致它具有极低的线性复杂度。同样,HV 运行时间也较短,并且运行时间也几乎是线性的。以 Modbus 为例,在分析 1 974 条消息时,NEMESYS 与 HV 都是在几秒内完成,而 Netzob 使用的时间是 HV 的 500 倍。

表 2 执行时间 s

协议	数据量	Netzob	NEMESYS	HV
S7COMM	334	8.770	0.630	1.251
	668	88.648	1.133	2.488
Modbus	1 974	1 528.349	1.138	3.154
	3 948	+∞	2.861	6.497
EGD	341	10.639	0.338	0.547
	674	67.533	0.556	1.032
DNP3.0	115	0.469	0.088	0.103

图 5 说明边界推断的质量只有在 Modbus 协议上是 Netzob>HV>NEMESYS;而 S7COMM,DNP3.0 和 EGD 协议上均为 HV>NEMESYS>Netzob。从表 2 可知,Netzob 的执行时间远远大于 HV 以及 NEMESYS。而后两者的执行时间相差无几。因此,综合字段划分质量和划分时间,HV 总体上是优于 Netzob 以及 NEMESYS 的,它有着较稳定的推断质量。HV 的执行时间几乎是线性的,当数据量较大时也能快速处理,而 Netzob 中的序列比对的复杂度是指数级别,当数据到达一定量就无法求解。

3 结束语

推断二进制协议的格式结构对于二进制协议分析十分重要。目前的协议逆向分析对于文本协议的研究较深,针对二进制协议进行逆向分析仍存在难点。字段划分是协议逆向过程中的前置步骤,协议逆向的准确度很大程度依赖于字段划分的质量。为解决上述问题,该文提出了一种新颖的较简单的未知二进制协议字段划分方法 HV。HV 首先单独分析每一条消息的内部结构;接着通过计算相邻消息之间的字节以及位翻转率进行字段划分;最后结合两次分段得到最终的字段划分结果。其他需要成对比较消息的方法复杂度在指数级别,HV 几乎只需要线性复杂度。并且与其它方案相比,此方案在推断字段边界的质量上也有着不错的表现。该文还定义了格式匹配分数来衡量字段划分的质量,相比传统的衡量指标,格式匹配分数更加适用于字段划分。

参考文献:

[1] TEAM C C. CAPEC-CAPEC-192:protocol reverse engineering (Version 2.6),Jun. 2014[EB/OL]. 2014. https://web. archive. org/web/20140725160124/http://capec. mitre. org,80.

[2] KRUEGER T, GASCON H, KRÄMER N, et al. Learning stateful models for network honeypots[C]//Proceedings of the 5th ACM workshop on security and artificial intelli-

gence. Raleigh:ACM,2012:37-48.

[3] LEITA C, MERMOUD K, DACIER M. Scriptgen: an automated script generation tool for honeyd[C]//21st annual computer security applications conference (ACSAC'05). Tucson:IEEE,2005.

[4] GASCON H, WRESSNEGGER C, YAMAGUCHI F, et al. Pulsar: stateful black-box fuzzing of proprietary network protocols[C]//Security and privacy in communication networks - 11th international conference, SecureComm 2015. Dallas:Springer,2015:330-347.

[5] WRESSNEGGER C, KELLNER A, RIECK K. Zoe: content-based anomaly detection for industrial control systems[C]// 2018 48th annual IEEE/IFIP international conference on dependable systems and networks (DSN). Luxembourg City: IEEE,2018:127-138.

[6] BEDDOE M A. Network protocol analysis using bioinformatics algorithms[J]. Toorcon,2004,26(6):1095-1098.

[7] WANG Y, YUN X, SHAFIQ M Z, et al. A semantics aware approach to automated reverse engineering unknown protocols[C]//2012 20th IEEE international conference on network protocols (ICNP). Austin:IEEE,2012:1-10.

[8] WANG L, JIANG T. On the complexity of multiple sequence alignment[J]. Journal of Computational Biology, 1994, 1(4):337-348.

[9] CUI W, KANNAN J, WANG H J. Discoverer: automatic protocol reverse engineering from network traces[C]//Proceedings of the 16th USENIX security symposium. Boston:USENIX,2007:1-14.

[10] BERMUDEZ I, TONGAONKAR A, ILIOFOTOU M, et al. Automatic protocol field inference for deeper protocol understanding[C]//Proceedings of the 14th IFIP networking conference, networking 2015. Toulouse:IEEE,2015:1-9.

[11] BERMUDEZ I, TONGAONKAR A, ILIOFOTOU M, et al. Towards automatic protocol field inference[J]. Computer Communications,2016,84:40-51.

[12] 张蔚瑶,张　磊,毛建瓴,等. 未知协议的逆向分析与自动化测试[J]. 计算机学报,2020,43(4):653-667.

[13] 黄　涛,付安民,季宇凯,等. 工控协议逆向分析技术研究与挑战[J]. 计算机研究与发展,2022,59(5):1015-1034.

[14] ZHANG Z, ZHANG Z, LEE P P C, et al. Toward unsupervised protocol feature word extraction[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(10): 1894-1906.

[15] SUN F, WANG S, ZHANG C, et al. Unsupervised field segmentation of unknown protocol messages[J]. Computer Communications,2019,146:121-130.

[16] WANG X, LV K, LI B. IPART: an automatic protocol reverse engineering tool based on global voting expert for industrial protocols[J]. International Journal of Parallel, Emergent and Distributed Systems,2020,35(3):376-395.

[17] JIANG D, LI C, MA L, et al. ABInfer: a novel field boundaries inference approach for protocol reverse engineering [C]//2020 IEEE 6th intl conference on big data security on cloud (BigDataSecurity), IEEE intl conference on high performance and smart computing (HPSC), and IEEE intl conference on intelligent data and security (IDS). Baltimore: IEEE,2020:19-23.

[18] 黎　敏,余顺争. 抗噪的未知应用层协议报文格式最佳分段方法[J]. 软件学报,2013,24(3):604-617.

[19] YU S Z. Hidden semi-Markov models[J]. Artificial Intelligence,2010,174(2):215-243.

[20] CAI J, LUO J Z, LEI F. Analyzing network protocols of application layer using hidden semi-Markov model[J]. Mathematical Problems in Engineering,2016,2016(pt.4):1-14.

[21] TAO S, YU H, LI Q. Bit-oriented format extraction approach for automatic binary protocol reverse engineering[J]. IET Communications,2016,10(6):709-716.

[22] KLEBER S, KOPP H, KARGL F. NEMESYS: network message syntax reverse engineering by analysis of the intrinsic structure of individual messages[C]//27th USENIX security symposium. Baltimore:USENIX,2018:1-13.

[23] MARCHETTI M, STABILI D. READ: reverse engineering of automotive data frames[J]. IEEE Transactions on Information Forensics and Security,2018,14(4):1083-1097.