doi:10.11959/j.issn.1000-436x.2016293

面向密码协议在线安全性的监测方法

朱玉娜[1,2]，韩继红[1]，袁霖[1]，范钰丹[1]，陈韩托[1]，谷文[1]
（1. 解放军信息工程大学三院，河南 郑州 450001；2. 解放军 91033 部队，山东 青岛 266035）

摘 要：为解决现有方法无法在线监测协议逻辑进行的低交互型攻击的问题，提出一种密码协议在线监测方法 CPOMA。首先构建面向密码协议的特征项本体框架，以统一描述不同类型的特征项，并基于该框架首次利用模糊子空间聚类方法进行特征加权，建立个体化的密码协议特征库；在此基础上给出自学习的密码协议识别与会话实例重构方法，进而在线监测协议异常会话。实验结果表明，CPOMA 不仅能够较好地识别已知协议、学习未知协议、重构会话，而且能够有效在线监测协议异常会话，提高密码协议在线运行的安全性。
关键词：密码协议识别；会话重构；在线安全性；本体；子空间聚类
中图分类号：TP393.08 **文献标识码**：A

Monitoring approach for online security of cryptographic protocol

ZHU Yu-na[1,2], HAN Ji-hong[1], YUAN Lin[1], FAN Yu-dan[1], CHEN Han-tuo[1], GU Wen[1]
(1. The Third College, PLA Information Engineering University, Zhengzhou 450001, China; 2. Troops 91033 of PLA, Qingdao 266035, China)

Abstract: Previous methods can not detect the low-interaction attacks of protocol logic. A cryptographic protocol online monitoring approach named CPOMA was presented. An ontology framework of cryptographic protocol features was constructed for the unified description of cryptographic protocol features with different types. Based on the framework, a feature weighting method was proposed by fuzzy subspace clustering first, and the individualized feature database of cryptographic protocols was built. On this basis, a self-learning method was presented for protocol identification and session rebuilding, and then abnormal protocol sessions were detected online. Experimental results show that CPOMA can identify protocols, rebuild sessions, detect abnormal sessions efficiently, and can improve the online security of cryptographic protocols.
Key words: cryptographic protocol identification, session rebuilding, online security, ontology, subspace clustering

1 引言

密码协议是互联网各种核心安全服务可靠运行的重要支撑，其安全性分析方法一直是信息安全领域的关键问题。传统方法通过形式化分析或自动化验证来监测协议自身缺陷，需要基于特定的攻击者模型和若干假设，只能给出理想情况下的安全性分析结果，对于协议运行过程中的某些动态因素往往不能准确判断。以经典的 SSL/TLS 协议为例，该协议发布的每个版本都有形式化方法证明安全，但随后又发现漏洞。由此，密码协议在线安全性分析技术已经成为新一代信息安全技术中亟待进行深入研究的关键问题。

密码协议运行过程中，频繁使用各种密码技术对关键信息进行加密和保护，其报文包含大量密文信息。攻击者无法解密密文，常常通过重放、转发密文对协议逻辑进行攻击。该种情况下，攻击过程仅表现为低交互性特征，不具备统计方面的特征；且与正常交互的报文具有相似的语法、语义特点。传统的入侵监测分析工具大多依靠流量分析或特定语义格式解析的手段进行监测，其中，基于流量分析的入侵监测主要针对高交互型攻击情况，如时间侧信道攻击，不适用于低交互型攻击；基于特定语义格式解析的入侵监测主要针对特定攻击模式，

收稿日期：2016-02-02；**修回日期**：2016-05-18
基金项目：国家自然科学基金资助项目（No.61309018）
Foundation Item: The National Natural Science Foundation of China (No.61309018)

如“心脏出血”攻击、版本回滚攻击等，不具备通用性，难以实现对未知攻击方法进行有效监测。因此现有方法均无法有效监测针对协议逻辑进行的低交互型攻击（如并行会话攻击、重放攻击）。

针对这一问题，本文提出了一种密码协议在线监测方法（CPOMA, cryptographic protocol online monitoring approach），实现了对密码协议低交互型攻击行为的语义级别的监测，有助于实现密码协议在线运行的安全性。主要贡献有：1) 构建面向密码协议的特征项本体框架，统一描述不同类型特征项；2) 首次基于模糊子空间聚类方法（FSC, fuzzy subspace cluster）进行特征加权，并建立协议个体化特征库；3) 针对未知密码协议，给出自学习的识别与会话实例重构方法；4) 实现密码协议在线监测平台，监测协议异常会话，为协议在线安全性分析提供支撑。

2 相关工作

2.1 密码协议识别与会话实例重构

要实现协议在线安全性分析，必须能够在线识别信息系统中报文数据的协议类型，重构协议会话实例，获取协议当前运行状态信息，进而有效监测当前协议是否存在安全隐患。因此，密码协议识别和会话实例重构技术是实现协议在线安全性分析的前提和基础。

现有密码协议识别方法主要有4个方面。1) 基于端口的方法：借助协议默认端口号识别，不适用使用动态端口的协议。2) 基于负载内容的方法[1,2]：通过匹配协议关键词识别，不适用于协议报文全加密的情况，且对新协议的识别具有滞后性。3) 基于流量统计特征的方法[3~8]。相同协议的网络流具有相似统计特征，可依据网络报文的流量统计特征识别协议。该类方法大都采用机器学习技术，能够识别全加密协议，但准确性和健壮性低于基于负载内容的方法。4)综合方法[9~12]。上述3类方法各有优缺点，一些研究试图将它们结合使用。PortLoad[9]结合基于端口的方法和基于负载内容的方法识别协议。TIE[10]支持以插件形式开发识别模块，实现多个识别方法的在线协同工作。NetraMark[11]结合11种不同的流量分类器，支持扩展新的识别方法并比较识别结果。文献[12]指出，硬编码实现的协议识别方法，每增加一种新的识别方法，修改识别规则都需要一个编写代码、重编译、系统重启的过程，无法在线升级新的识别方法和频繁改变的识别规则；并针对该问题，结合各种识别方法提出一种可扩展的识别架构，但无法识别未知协议。

由上可知，结合各种方法的优点，建立统一的协议识别框架是协议识别领域的重要研究方向。该框架需要支持协议特征库的更新和扩展，并且有效解决未知协议的识别问题。

协议会话实例重构需要进一步识别协议报文类型，并重构报文关键项的语法、语义和交互步骤，这需要以协议的格式描述信息为基础。现有协议分析工具（如Wireshark）仅能重构已知规范协议会话，无法恢复未知规范协议会话。因此，需要自动解析未知密码协议格式信息。基于网络报文流量信息的方法[13,14]仅考虑报文载荷中的明文信息，不适用于包含大量密文信息的密码协议。为此，朱玉娜等[15]提出一种新的面向未知密码协议的格式解析方法SPFPA，解析可用明文格式特征，并挖掘协议报文中包含的密文数据特征。

2.2 密码协议异常会话监测

为保障密码协议运行过程中的安全性，研究人员基于流量分析[16~19]或特定语义格式解析（如Snort心脏出血漏洞下行监测规则）监测密码协议异常。ProtoMon[16]通过监测协议执行进程监测协议会话的安全性；文献[17]采集分析多种形式的元数据，并基于统计和模式识别方法监测协议的异常状态；文献[18]提出了一种非参数累积和算法追踪可能的攻击者；文献[19]给出了基于统计模型的网络数据报文监测方法。

上述方法均在攻击报文与正常报文有明显区别或者高交互型攻击下监测协议异常。对并行会话攻击等低交互型攻击而言，攻击过程与正常交互时的协议报文语法格式相同，无法利用上述方法有效监测。文献[20]为每类密码协议建立运行状态转换规则库和已知攻击特征库，结合密码协议参与者的活动信息与特征库的映射关系，监测密码协议运行行为，但主要监测密码协议的已知逻辑漏洞，不能监测未知攻击，不具有通用性。为此，针对协议逻辑的低交互型攻击，需要给出新的具有通用性的协议安全性监测方法。

3 CPOMA总体框架

密码协议在线监测方法CPOMA的总体框架如图1所示。

CPOMA具体各模块功能描述如下。

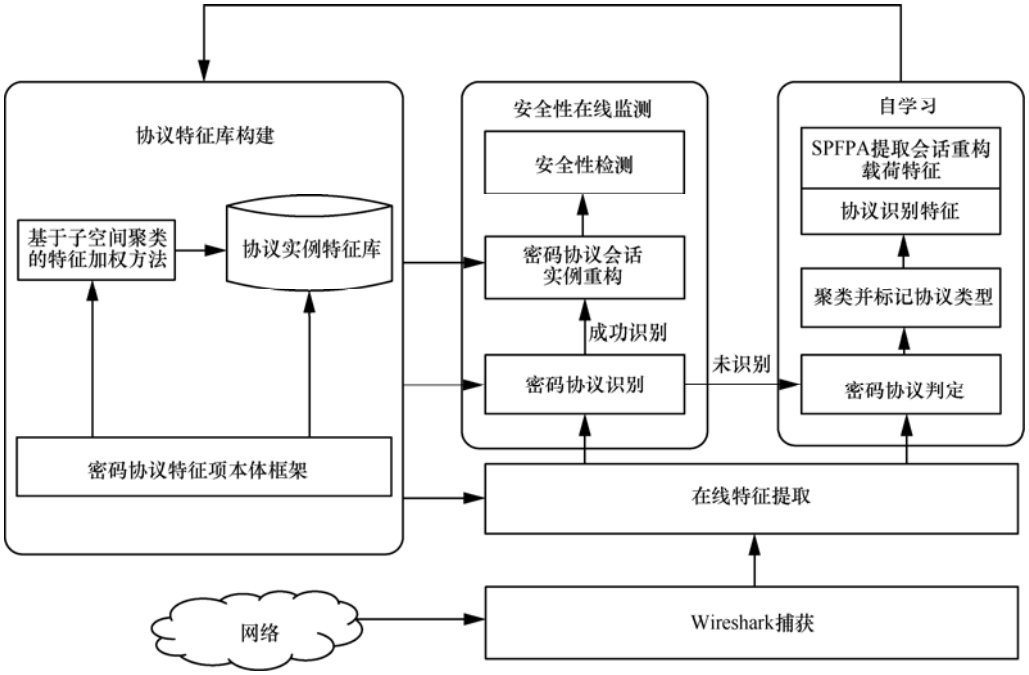


图 1　CPOMA 总体框架

1) 协议特征库构建

分析密码协议识别特征项和会话实例重构特征项，并基于 Methontology 方法构建密码协议特征项本体框架。考虑每一维特征对不同类别的识别贡献程度，基于 FSC 方法为每类协议的各个识别特征项获取相应的权重系数。基于特征项本体框架和获取的加权特征为每类协议分别给出相应本体实例，构建个体化的协议实例特征库。

2) 安全性在线监测

基于特征项描述框架在线提取实际网络环境下获取的报文特征，并根据协议实例特征库的识别规则进行推理，识别协议类型。基于识别结果和密码协议实例特征库，进一步识别报文类型，恢复报文关键项的语法、语义、交互步骤，重构协议会话实例。监测协议当前状态，若发现攻击行为，则及时阻断协议会话，避免攻击者进一步攻击。

3) 自学习

依据协议的规范性和报文中密文数据的随机性，启发式判定未识别报文是否属于密码协议。对未知密码协议聚类并标记其类型，提取协议特征项存入协议实例特征库，用于后续监测。

4　协议实例特征库构建

4.1　面向密码协议的特征项本体框架

本节分析密码协议的识别特征项和会话实例重构特征项，在此基础上构建密码协议特征项本体框架。

4.1.1　密码协议特征项分析

密码协议识别特征项主要有以下 3 类。1) 端口特征。2) 数据报文载荷特征：①关键词特征，关键词是在报文格式中用于标识协议报文类型和传递相关控制信息的协议字段，如协议名称、版本号、命令码、标识信息等，绝大多数的网络协议都会在报文格式中定义一个或多个关键词，关键词在协议中频繁出现，是组成协议特征的重要元素；②负载统计特征，将负载的前 N byte 作为特征矢量以识别协议，Haffner[2]指出只需要流的前 64 byte 负载就可以挖掘协议识别特征。3) 流量统计特征：Moore 等[3]给出了网络流中用于协议识别的 249 种属性特征，其中，数据分组大小、分组到达时间间隔是关键特征[6,8]。

本文采用端口、流中前 64 byte 负载的关键词序列和字节分布、流中前 N 个数据分组的大小和分组到达时间间隔作为密码协议识别特征。端口和负载关键词为精确特征。而同一协议的统计特征（负载字节分布、数据分组大小、分组到达时间间隔）并非严格一致，通常借助于机器学习方法完成识别。现有基于统计特征的识别方法大都为不同协议选择统一的特征集合，而每一维特征对不同类别可能具有不同的贡献，应引入特征权重来强化重要特征的积极作用，削减冗余特征的不利影响。

会话实例重构需要识别协议报文类型，并重构

报文关键项的语法、语义和交互步骤。本文密码协议的会话实例重构特征项主要有以下几方面。1) 报文类型识别特征：协议会话由不同类型的报文组成，每类报文实现特定的功能，具有不同的语法语义，可根据协议会话过程中的数据报文载荷特征和流量统计特征识别报文类型。2) 报文关键项重构特征：①关键词特征，重构会话不仅需要考虑与协议识别相关的前 64 byte 关键词，还需要考虑完整协议会话过程中的其他关键词；②密文数据特征，攻击者常常通过重放、转发密文攻击协议，应充分利用密文数据特征。3）协议时序行为特征：协议时序行为体现了协议状态之间的转换关系。一次协议会话可看作是协议的一种状态转换路径。

4.1.2 协议特征项本体框架

不同类型的密码协议特征项在格式、取值范围等方面存在巨大差异。本体是指共享概念模型的明确的形式化规范说明，用于描述概念以及概念之间的关系，支持逻辑推理、便于知识重用，应用广泛。本文基于本体描述协议特征项，建立可扩展的协议特征项描述框架。一方面，该框架可以随时修改其结构、功能，方便增加规则，支持特征库更新扩展，便于重用和共享；另一方面，可以基于该框架实例化特定类型协议，为每类协议分别建立相应的特征和识别规则，构建协议个体化特征库。

一个本体结构是一个 5 元组 $O:=\{C,R,H^C,rel,A^O\}$ [21]。其中，C 为概念集合，R 为关系集合，H^C 表示概念层次，rel 表示概念间的非分类关系，A^O 为使用某种逻辑语言表达的本体公理集合。Methontology 方法与其他本体构建方法相比，更符合 IEEE1074-1995 软件开发标准，具有更强的表示知识的能力。本文基于 Methontology 方法建立协议特征项本体框架，描述协议特征项基本信息，如图 2 所示。

密码协议识别特征项、会话实例重构特征项以及特征项相关的属性概念统称为协议概念 C_p。协议识别和会话实例重构将流作为一个整体考虑，因此在密码协议特征项本体框架中将流 Flow 作为根概念。Flow 包括 2 个属性概念：特征项 Signature 和协议类型 ProtocolType，其中，ProtocolType 可由

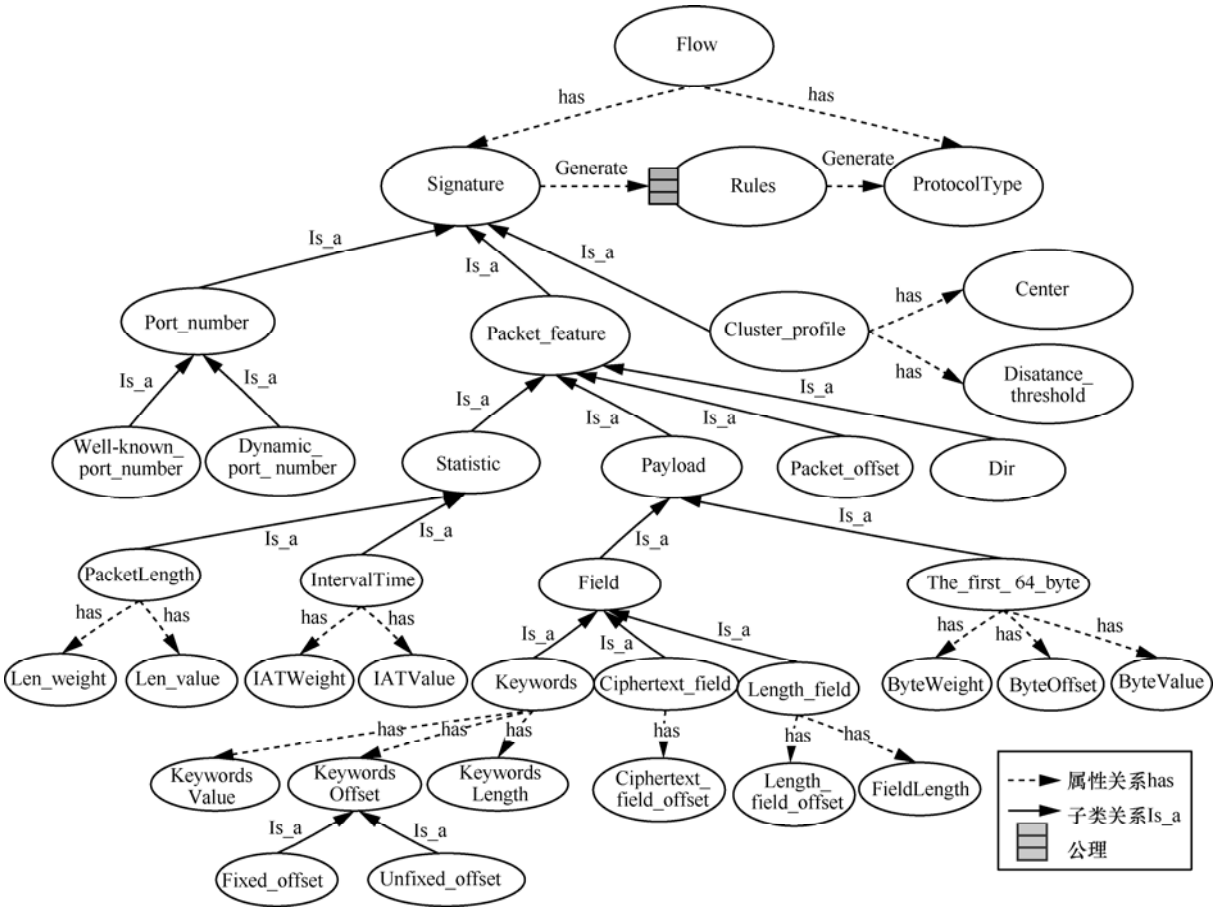


图 2 密码协议特征项本体框架

Signature 生成的公理得到。Signature 包括如下几项。1）端口特征 Port_number。2）数据分组特征 Packet_feature，包括：①数据分组统计特征 Statistic——数据分组大小 PacketLength 和分组到达时间间隔 IntervalTime；②载荷特征 Payload：域 Field（关键词 Keywords、密文域 Ciphertext_field、长度域 Length_field）和负载统计特征 The_first_64_byte；对流中第 1 个数据分组而言，若其负载大于等于 64 byte，则载荷特征直接从该数据分组中获取；若其负载小于 64 byte，则还需从后续数据分组中进一步获取与流中前 64 byte 相关的载荷特征；③时序行为相关的特征——数据分组偏移 Packet_offset 和数据分组方向 Dir。3）统计特征相关的类簇特征 Cluster_profile，包括类簇中心 Center 和距离阈值 Distance_threshold。随后，对特征项概念根据密码协议的运行特点进一步确定相关的属性概念，如 PacketLength 存在 2 个属性概念：权重系数 Len_weight 和值 Len_value。

协议概念关系 R_{P} 包括子类关系和属性关系，记为 $R_{\mathrm{P}}=\{\mathrm{Is\_a}, \mathrm{has}\}$。$rel_{\mathrm{P}}: R_{\mathrm{P}} \to C_{\mathrm{P}} \times C_{\mathrm{P}}$ 表示 2 个概念之间的某种关系。记 $C_1, C_2 \in C_{\mathrm{P}}$ 为 2 个概念，has(C_1,C_2)表示 C_2 是 C_1 的属性；Is_a (C_1,C_2)表示 C_2 是 C_1 的子类。C_{P} 中各种协议概念及它们之间的关系构成协议概念层次 H_{P}^{C}。

SWRL(semantic Web rule language)是由以语义的方式呈现规则的一种语言，目前已成为语义 Web 逻辑层的标准语言。SQWRL(semantic query-enhanced web rule language)是 SWRL 的扩展语言。本文基于 SQWRL 描述协议特征项本体框架中的协议识别规则 A_{P}^{O}。1) 端口识别规则和数据分组载荷识别规则：端口特征和数据分组载荷的关键词特征为精确特征，对其进行匹配即可识别密码协议。端口识别规则如下：Flow(?x) ∧Port_number(?x, ?y) ∧swrlb:equal(?y, value)→ ProtocolType；数据分组载荷识别规则如下：Flow(?x)∧KeywordsValue(?x,?y) ∧KeywordsOffset(?y,?z) ∧swrlb:equal(?y, value$_1$) ∧swrlb:equal(?z, value$_2$)→ProtocolType。2) 流量统计特征识别规则：同一协议的负载统计特征和流量统计特征并不完全一致，大都借助于机器学习进行识别。对协议样本进行聚类，获取每类协议的类簇中心、类簇划分、特征权重系数，并为每个类簇设定距离阈值。计算待识别样本与每个类簇中心的距离（由于引入特征权重，分别计算样本特征向量 $\boldsymbol{x}''=(x''_1, x''_2 \mathrm{L}\ , x''_d)$ 与所有协议类簇中心 z_j 之间的加权欧式距离 $d(\boldsymbol{x}''-z_j)$），若距离均大于给定的距离阈值，则为未知协议，即(Flow(?x)∧Center(?z_j)∧Distance_threhold(?r_j) ∧ swrlb:greaterthan (?d(x−z_j), r_j)) →UnknownProtocolType；否则选择距离最近的类簇中心作为对应的协议类型。即 Flow(?x) ∧Center(?z_j) ∧Distance_threhold(?r_j) ∧swrlb:lessthan (?d(x−z_j), r_j))∧sqwrl:min(?d(x−z_j))→ ProtocolType。

4.2　基于 FSC 的特征加权方法

子空间聚类方法可以在对数据样本聚类划分的过程中，得到各个数据簇对应的特征子集，目前已成功应用于文本分类等领域。SubFlow[8]首次采用硬子空间聚类方法构建协议个体化特征库，提高了识别的准确性和健壮性，但在提取协议特征阶段，要求流量由同一类协议组成，不适用多种协议混杂的情况。与硬子空间聚类方法对比，软子空间聚类具有更好的适应性和灵活性，已成为学术界的研究热点，FSC[22]是其中的典型方法。

为提高基于统计特征的识别效果，本文基于 FSC 获取各个特征的权重系数，为不同类型协议构建不同的加权统计特征。

4.2.1　FSC 方法

FSC 输入为数据集 $D=\{x_1, x_2, \mathrm{L}\ , x_n\}$、类簇数目 k、模糊加权指数 α。输出为类簇 U、类簇中心矩阵 $\boldsymbol{Z}$、特征加权系数矩阵 $\boldsymbol{W}$。其中，$\boldsymbol{Z}=\left[z_1, z_2, \cdots, z_k\right]^{\mathrm{T}}$，$z_j$ 为第 j 类的中心。$\boldsymbol{W}=\{w_{jh} \mid 1 \leqslant j \leqslant k, 1 \leqslant h \leqslant d\}$，$w_{jh}$ 表示第 h 个特征对于第 j 个数据簇的重要性，$0 \leqslant w_{jh} \leqslant 1$，$\sum_{h=1}^{d} w_{jh}=1$。对于给定的数据集 D，FSC 在聚类过程中考虑各个特征的识别贡献程度，在得到各个类簇中心 $\boldsymbol{Z}$ 的同时，还为每个类簇获得各个特征的权重分配。

4.2.2　特征加权方法

1) FSC 初始点选择算法

FSC 方法对初始点敏感，初始中心选择不当容易陷入局部最优解。为此，本文借鉴 $k$means++的思想，选取彼此距离尽可能远的样本点作为初始类簇中心。从数据集 D 中随机选择 1 个点作为第 1 个类簇中心，并依据规则 $x \mid \max_{x \in X}(\min_{v \in V}(d(x, z))$ 选取其他类簇中心。

2) 特征加权

特征加权流程如图 3 所示。

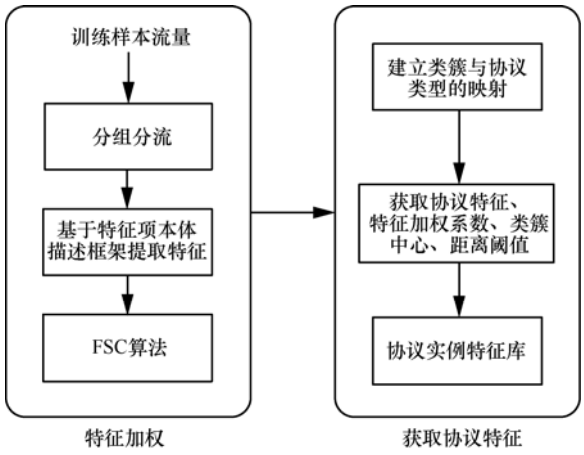


图 3 基于 FSC 的特征加权

Step1 基于 FSC 方法获取各个特征的权重系数。首先利用初始点选择算法从协议数据集 D 中选择 k 个中心点，随后执行 FSC 方法，得到 D 的类簇划分 U、类簇中心矩阵 $\boldsymbol{Z}$ 和特征加权系数矩阵 $\boldsymbol{W}$。

Step2 建立类簇与协议类型的映射，并获取协议特征。

FSC 聚类结果是未知类别标签的类簇集合，需要进一步明确每个类簇对应的协议类型。记协议类型集合为 $L=\{L_1,\mathrm{L}\ ,L_m\}$，$N_j^s$ 为类簇 z_j 中协议类型 $L_s(1\leqslant s\leqslant m)$的样本数目；$N_j$ 是类簇 z_j 的样本数目。由最大似然估计可知 $p(l=L_s\mid z_j)=\dfrac{N_j^s}{N_j}$，因此类簇与协议类型的映射函数为 $l(z_j)=\arg\max p(l=L_s\mid z_j)=\arg\max\left(\dfrac{N_j^s}{N_j}\right)$。即对每一个类簇中的样本标记协议类型，类簇中样本数目最多的类型标签记为该类簇的协议类型。

记第 j 个类簇的特征加权矢量为 $\boldsymbol{w}_j=[w_{j1},w_{j2},\mathrm{L}\ ,w_{jd}]$，类簇中心为 z_j，x' 为该类簇的某一样本点，则 x' 与 z_j 的加权欧式距离为 $d(x'-z_j)=\sqrt{\sum_{h=1}^{d}w_{jh}\left(x'_h-z_{jh}\right)^2}$。与文献[4]类似，基于类簇内距离的方差对每一个协议类簇设定距离阈值 $r_j=T\sigma_j^2$，其中，σ_j^2 为第 j 个类簇内加权欧式距离的方差，用于描述类簇内距离的离散程度，T 为正值参数，用于调整方差对 r_j 的影响程度。

在确定类簇对应的协议类型后，该类簇对应的类簇中心、特征、特征加权系数、距离阈值即为相应的协议特征。

4.3 密码协议实例特征库

根据协议特征项本体描述框架，对特定类型协议分别进行实例化，构建协议实例特征库。1）确定协议特征项各个属性的值。对端口特征，根据 well-known 注册端口号进行实例化。对数据报文载荷特征，已知规范协议根据其规范进行实例化，未知规范协议则依据 SPFPA 方法[15]进行逆向，提取协议载荷特征；对流量统计特征，基于 FSC 方法进行加权，将获取的加权统计特征存入协议本体特征库。2）描述协议识别规则。以 Https 协议端口识别规则为例，在 well-known 端口号（443）上运行的都为固定类型 Https 的协议，识别规则为 Flow(?x) ∧ port_number(?x,?y) ∧ swrlb:equal(?y, 443)→ ProtocolType(?x, https)。

Protégé是由斯坦福大学开发的开源本体编辑器。由于其开放性和兼容性，Protégé成为目前本体编辑的首选工具，应用广泛。本文采用 Protégé编辑协议本体库，SWRL 规则通过 Protégé的 SWRL Tab 插件编写。

5 安全性在线监测

5.1 密码协议识别与会话实例重构

基于特征项本体描述框架在线提取待识别样本流的协议特征向量，并基于本体库中的识别规则进行推理，输出协议类型。

随后基于识别结果重构协议会话实例。按照参与方数目，可将密码协议划分为两方密码协议和多方密码协议。两方密码协议一次会话过程包含在一个 TCP 连接或一个 UDP 双向流中。多方密码协议会话则分布在多个单向流中。现有方法不能确定哪些流的报文具有关联关系，无法恢复多方密码协议会话过程。本文针对这一问题，提取属于同一次会话的流之间的关联特征，确定属于同一次会话的流[23]。

在此基础上，结合密码协议特征描述框架的目标密码协议消息相关特征（数据报文大小、方向、偏移位置、载荷等），即可确定每个会话中的报文类型，从而对目标协议一次完整的消息交互序列进行构建，并构建关键项的语法、语义、交互步骤。

5.2 安全性监测

本文基于如下方法实现对协议实现逻辑的低交互性攻击监测。

Step1 并行会话监测。攻击者无法解密密文，

通过转发、重放密文进行攻击，因此并行会话的出现是攻击发生的必要条件。可根据会话重实例构结果监测并行会话。

Step2　重放项检查。在监测到并行会话序列后，进一步对相关会话中关键消息项进行重放项检查。本文采用随机抽取节点的方法对密文进行比对。N 次随机抽取位置如果都一致的情况下（N 的取值与密文长度相关），可判定为 2 个密文项相同。

Step3　攻击判定。由密文的随机性可知，正常交互情况下不同会话中的报文密文项不同。若发现密文部分存在重复内容，则存在攻击。

Step4　报警及攻击定位。攻击发生后通知用户，锁定攻击方 IP 地址，并存储相关消息。

6　自学习反馈机制

当出现未知密码协议时，协议识别效果下降。本文引入自学习机制，获取未知密码协议特征，用与后续识别与会话实例重构。

Step1　启发式判断未识别流是否为密码协议。

1) 根据长度变化范围区分协议和传输流量

协议具有特定的规范，流中前几个分组长度分布在一定取值范围内，一般变化较大。传输数据时，通常需要在 IP 层根据 MTU 对数据进行分片，流中分组长度大部分为固定值。由此依据长度变化范围区分协议和传输流量。若第一个分组小于 MTU，执行 2)。若大于 MTU，由于协议可能存在较长的证书数据（证书一般需要 2~3 个数据分组），进一步进行处理，若第 2~3 个数据分组长度等于 MTU，则对该流不再进行处理，否则，执行 2)。

2) 依据密码协议密文数据的随机性区分密码协议和非密码协议

①判断第 1 个数据分组前 32 byte 的密文随机性，对全密文数据而言，对其前 32 byte 进行熵估计，即可判断该数据是否加密[24]。若是，执行 Step2，否则执行②。

②对密文而言，连续 5 byte 第一个比特值同时为 0 或者同时为 1 的概率为$\left(\frac{1}{2}\right)^5 = 0.031\,25 < 0.05$，为小概率事件。对文本协议而言，其明文为 ASCII 码，取值范围为 0~127，字节第一个比特值为 0，明文中经常出现 5 个连续 ASCII 码。对于二进制协议，由于协议规范具有特定语义，也经常出现 5 个连续字节第一个比特值为 0 的情况。可利用连续 5 byte 第 1 个比特是否相同大致判断是否存在密文。对第 1 个数据分组载荷部分查找可能的密文区间。首先对载荷字节进行编码，连续 5 byte 第一个比特相同则编码为 0，不同则编码为 1。鉴于密文的随机性，密文数据编码后出现 0 为小概率事件。协议密文一般长度为 16~512 byte，本文对编码为 1 的负载字节区域，设置 16 byte 的滑动窗口，该窗口始终包含该负载字节区域，并对该窗口包含的字节区域进行随机的频数测试。若存在某滑动窗口且该窗口包含的字节区域能够通过频数测试，则数据分组中可能存在密文域，该流可能为密码协议类型，执行 Step2，否则执行③。

③密码协议流中最后 1 个数据分组通常包含密文数据。对流中最后 1 个数据分组执行与第 1 个数据分组相同的步骤，若判定该分组可能存在密文域，执行 Step2；否则，对该流不做处理。

Step2　聚类并标记协议类型。对可能为密码协议的流，基于特征项描述框架在线提取特征，提取流第一个数据分组前 64 byte 载荷数据和前 N 个数据分组长度、分组间隔时间。当未识别密码协议的流达到一定数目后，通过 FSC 方法进行聚类，并对类簇得到的协议簇进行标记。其中，在选择 FSC 初始点时，已存在的类簇中心作为初始点，以降低复杂度。

Step3　提取协议特征，用于后续识别和会话重构。将协议类簇对应的识别特征存入协议实例特征库，参与后续识别过程；对同一类型协议依据笔者提出的 SPFPA 方法[15]进行逆向，首先基于序列模式挖掘方法提取协议的关键词序列特征，并在此基础上利用密文数据的随机性特征确定密文域，充分利用协议密文数据特征，从而有效解析协议格式，提取协议载荷特征，存入协议实例特征库，用于后续会话实例重构；未聚类的未识别密码协议继续参与后续聚类。

7　密码协议在线监测平台

基于 CPOMA 方法，本文设计了面向密码协议的在线监测平台，平台效果如图 4 所示。该平台包含报文识别模块、会话重构模块和攻击监测模块，能够实现在线监测针对协议逻辑的攻击行为。

7.1　实验环境

基于 CPOMA 平台，选取 SSL 协议、SSH 协议、

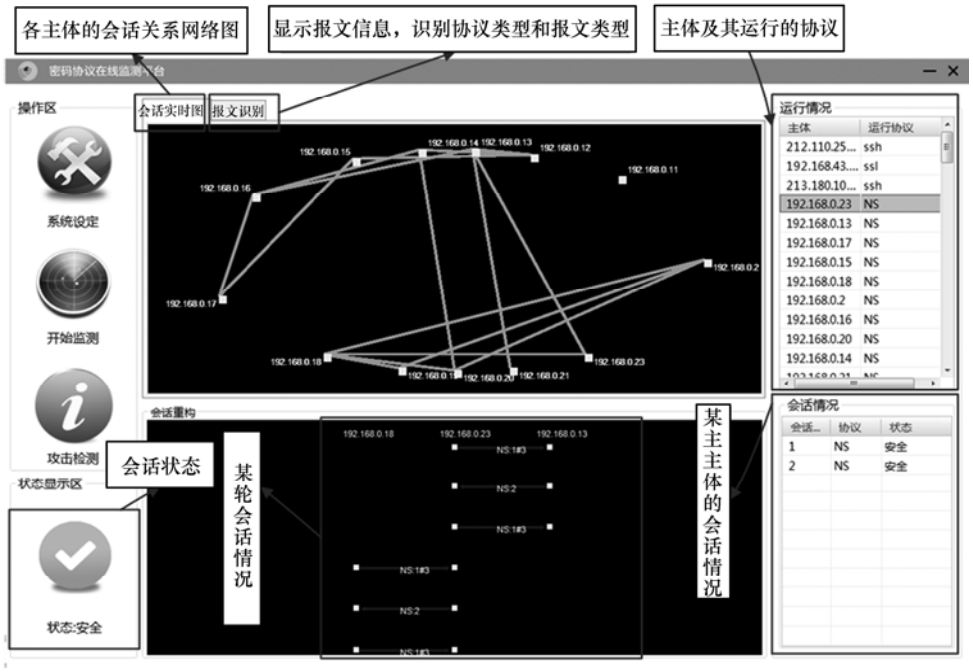


图 4　密码协议在线监测平台

NS 协议以及 Skype 协议进行实验，其中，SSL、SSH 和 Skype 协议是网络中广泛应用的密码协议；NS 公钥协议属于经典基础密码协议。

协议流量数据集如表 1 所示。第 1 部分为包含 SSL 协议的广域网流量。第 2 部分来源于 InfoVisContest 数据集[注1]，为包含 SSH 协议的网络流量。第 3 部分由实验室局域网环境产生，为包含 NS 公钥协议的网络流量，包括正常协议流量和并行会话攻击的协议流量，其中，NS 公钥协议的应用程序采用 Spi2Java 工具生成，并在各个主机上运行。第 4 部分来源于 Tstat 数据集[注2]，为 Skype 的 UDP 流量。第 5 部分来源于广域网，为普通网络通信协议（Http 协议、FTP 协议）流量。

表 1　　协议数据集

协议	流数量	数据来源
SSL	2 950	WAN
SSH	270 548	InfoVisContest
NS	2 000	LAN
Skype	2 000	Tstat
http	2 000	WAN
FTP	2 000	WAN

实验主要验证 CPOMA 的识别效果、自学习效果、会话重构效果以及异常会话监测效果。Skype 协议是私有协议，其协议规范不公开，无法确定会话重构效果，本文将其用于识别效果和自学习效果的验证。

首先，基于 Lua 脚本对数据集进行处理，获取与协议相关的信息（如通信双方 IP、端口、载荷内容、分组长度等）。对协议识别特征通过 Z-score 进行归一化处理，使数据的各维特征都在[0,1]。从 SSL、SSH 和 NS 中分别提取 500 个完整会话，作为训练集，用于构建协议实例特征库；数据集其余部分作为测试集，采用 JESS 推理引擎进行知识推理，识别协议，并基于实例特征库进一步解析协议，重构会话并监测是否存在异常。

7.2　参数设置

聚类参数主要有 FSC 模糊指数 α、FSC 聚类数目 k 以及数据分组个数 N_{p}。文献[11]建议 α 设置在 2 附近，与文献[22]相同，设置 $\alpha=2.1$。聚类数目 k 则采用标准化互信息(NMI, normarlized mutual information)进行设置。$NMI(X,Y)=2\dfrac{I(X;Y)}{H(X)+H(Y)}$。其中，$X$ 和 Y 分别为样本集中表示协议类型和类簇标识的变量，$I(X;Y)=\sum\limits_{y\in Y}\sum\limits_{x\in X}p(x,y)\log\left(\dfrac{p(x,y)}{p(x)p(y)}\right)$ 为 X 和 Y 的互信息，$H(X)$ 和 $H(Y)$ 分别为 X 和 Y 的熵。$NMI(X,Y)$ 取值为 0~1，值越大，聚类效果越好。当协议类型

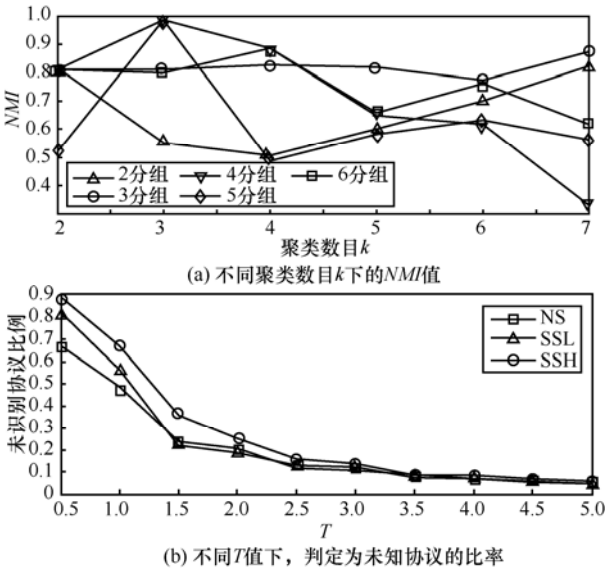
注1 http://2009.hack.lu/index.php/InfoVisContest。

注2 http://tstat.tlc.polito.it/traces-skype.shtml。

与类簇标识一一对应时，$NMI(X,Y)=1$。

不同聚类数目下的 *NMI* 值如图 5(a)所示。训练集中存在 3 类协议——NS 、SSL、SSH。当 $k=3$，$N_p=4$ 或 $N_p=5$ 时，*NMI* 为最大值。由于设定 $N_p=4$ 与设定 $N_p=5$ 相比，CPOMA 复杂度更低，效率更高，本文设定 $k=3$，$N_p=4$。

识别参数为距离阈值 r_j 中的 T，依据文献[4]方法进行设定。对训练集在不同 T 值情况下进行识别，并统计识别为未知协议类型的比例，如图 5(b)所示。当 T 增大时，判定为未知协议类型的样本数目减少，识别新协议类型的概率也随着下降。将可以较好识别已知协议的最小 T 值作为阈值，设定 $T=2.5$。



(a) 不同聚类数目k下的NMI值

(b) 不同T值下，判定为未知协议的比率

图 5　实验参数设置

7.3　实验结果

1) 协议识别结果

采用如下性能指标衡量识别效果。记测试集中某协议 A 的样本数目为 N。N_1 表示被正确识别为 A 的样本数，N_2 表示非 A 被错误识别为 A 的样本数，识别率=$\frac{N_1}{N}$，误识别率=$\frac{N_2}{N_1+N_2}$。识别率越高，误识别率越低，相应的识别效果越好。

在不加入自学习反馈机制的情况下，分别采用 *k*-means 和 FSC 方法获取协议特征并识别协议，由于子空间聚类考虑了不同协议特征的权重，识别效果相对 *k*-means 方法更好，如图 6 所示。

2) 会话重构效果

对识别的 SSL 协议、SSH 协议和 NS 协议进行会话重构。SPFPA 方法[17]根据识别结果可以较好地解析协议。多方密码协议会话识别方法[18]可以较好地构建密码协议会话流，在此基础上能够进一步识别报文类型，重构协议会话实例。采用会话重构率指标评价会话重构效果。记测试集中某协议 A 的会话样本数目为 N'，N_1' 表示被成功重构的会话数目，协议 A 的会话重构率为 $\frac{N_1'}{N'}$。不同训练样本数的协议会话重构率如图 7 所示，当训练集中某协议的会话数目 M 大于 100 时，会话重构率在 92.3%以上。

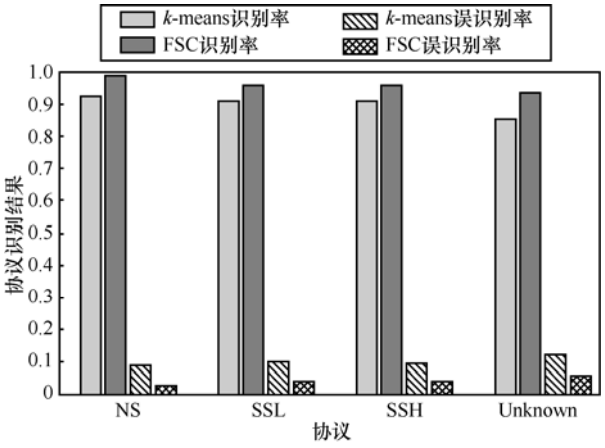


图 6　协议识别结果

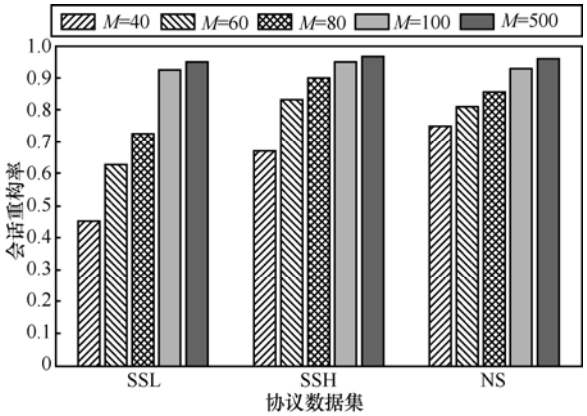


图 7　训练集不同大小时的协议会话重构率

3) 自学习反馈效果

在加入自学习反馈机制的情况下，采用 FSC 方法识别协议。依次选择 4 个测试集进行验证，每类包含 500 个协议会话。第 1 部分为包含 SSL、SSH、NS 的流量，第 2 部分在第 1 部分基础上增加 Http 和 FTP 协议，第 3 部分在第 1 部分基础上增加 Skype 协议流量、密文数据传输流量。第 4 部分包括所有

表 2　　自学习结果

测试集	流量	识别比例	自学习结果
1	SSL、SSH、NS	95.43%	较好识别 SSL、SSH、NS 协议
2	SSL、SSH、NS、Http、FTP	57.2%	Http、FTP 判定为非密码协议
3	SSL、SSH、NS、Skype、加密数据传输	56.4%	Skype 判定为未知密码协议，加密数据传输判定为非密码协议
4	SSL、SSH、NS、Http、FTP、Skype	63.9%	Skype 协议识别率为 97.2%

协议。记测试集中协议样本数目为 M，M_1 表示成功识别的协议样本，记识别比例为 $\frac{M_1}{M}$。结果如表 2 所示，对测试集 2 进行自学习后，Http 协议和 FTP 协议判定为非密码协议；对测试集 3 进行自学习后，可以成功将 Skpye 协议判定为未知密码协议流量，并形成新簇，提取其识别特征，加密数据传输流量则判定为非加密流量；在测试集 4 中能够识别 Skype 协议，其识别率为 97.2%。

4) 异常会话监测效果

测试集的 NS 协议流量中存在 1 397 次正常的密码协议会话，103 次并行会话攻击。CPOMA 成功监测 98 次攻击，攻击监测正确率为 100%，漏报率为 6.7%。攻击监测效果如图 8 所示，针对监测到的攻击，可提供实时报警，定位攻击者。

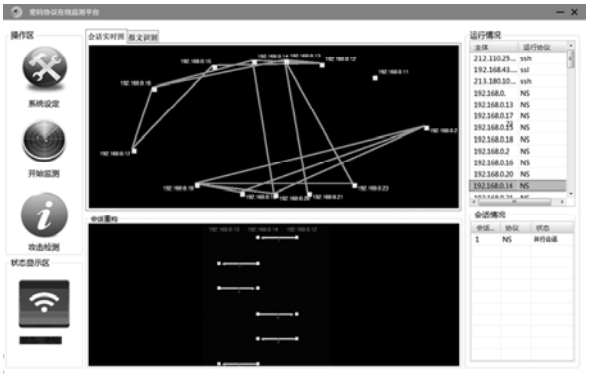


图 8　NS 异常会话监测效果

8　结束语

本文提出了一种密码协议在线监测方法 CPOMA。该方法建立了特征项本体描述框架，给出了基于 FSC 的特征加权方法，并构建协议特征库。在此基础上进行协议识别和会话实例重构，进而监测协议异常会话。实验结果表明，该方法能够较好地监测协议会话，为协议动态安全性分析提供支撑，但 CPOMA 还存在一定的局限性。1）目前，多方密码协议的识别主要针对可以获取同类多方密码协议流量的情况，需要提前进行训练并提取多方密码协议特征，下一步提出多方密码协议的自学习识别方法。2）FSC 对初始点选择敏感、容易陷入局部最优解，出现若干类簇合并的现象，需要结合半监督学习和 FSC，进一步提高协议识别正确率。

参考文献：

[1] BERNAILLE L, TEIXEIRA R. Early recognition of encrypted applications[C]//The 8th International Conference on Passive and Active Network Measurement. Belgium, c2007: 165-175.

[2] HAFFNER P, SEN S, SPATSCHECKO, et al. ACAS: automated construction of application signatures[C]//ACM SIGCOMM Workshop on Mining Network Data. Philadelphia, PA, USA, c2005: 197-202.

[3] MOORE A, ZUEV D, CROGAN M. Discriminators for use in flow-based classification: technical report, RR-05-13[R]. UK: Quecn Mayr University of London, 2005.

[4] BERNAILLE L, TEIXEIRA R，SALAMATIAN K. Early application identification[C]//ACM CoNEXT, Lisboa, Portugal, c2006.

[5] ZHANG J, XIANG Y, WANG Y, et al. Network traffic classification using correlation information[J]. IEEE Transactions on Parallel & Distributed Systems, 2013, 24(1): 104-117.

[6] BARALIS E M，MELLIA M，GRIMAUDO L. Self-learning classifier for internet traffic[J]. IEEE INFOCOM, Turin, Italy, c2013, 11(2): 423-428.

[7] DIVAKARAN D M，SU L，LIAU Y S, et al. SLIC: self-learning intelligent classifier for network traffic[J]. Computer Networks, 2015, 91: 283-297.

[8] XIE G W, ILIOFOTOU M, KERALAPURA R, et al. SubFlow: Towards practical flow-level traffic classification[C]//IEEE INFOCOM. Orlando, Florida, USA, c2012: 2541-2545.

[9] ACETO G, DAINOTTI A, DONATO W, et al. PortLoad: taking the best of two worlds in traffic classification[C]//IEEE INFOCOM. San Diego, 2010:1-5.

[10] DONATO WD，PESCAPÈ A，DAINOTTI A. TIE: a community-oriented traffic classification platform[C]//International Workshop on Traffic Monitoring and Analysis (TMA), Springer Berlin Heidelberg. c2009.

[11] LEE S, KIM H-C, BARMAN D, et al. NeTraMark: a network traffic classification benchmark[C]//ACM SIGCOMM. Toronto,ON, Canada, c2011.

[12] 张众, 杨建华, 谢高岗. 高效可扩展的应用层流量识别架构[J]. 通信学报, 2008, 29(12): 22-31.
ZHANG Z, YANG J H, XIE G G. Efficient and extensible architecture of traffic identification at application layer[J]. Journal on Communications, 2008, 29(12): 22-31.

[13] BEDDOE M. The Protocol information project[EB/OL]. http://www.tphi.net/ awalters/ PI.html.

[14] CUI W D, KANNAN J, WANG H J. Discoverer: automatic protocol reverse engineering from network traces[C]//The 16th USENIX Security Symposium on USENIX Security Symposium. Berkeley: USENIX, c2007: 199-212.
[15] 朱玉娜, 韩继红, 袁霖, 等. SPFPA：一种面向未知密码协议的格式解析方法[J]. 计算机研究与发展, 2015, 52(10): 2200-2211.
ZHU Y N, HAN J H, YUAN L, et al. SPFPA: a format parsing approach for unknown security protocols[J]. Journal of Computer Research and Development, 2015, 52(10): 2200-2211.
[16] JOGLEKAR S P, TATE S R. Protomon: embedded monitors for cryptographic protocol intrusion detection and prevention[C]// International Conference on Information Technology: Coding and Computing, 2004. ITCC 2004. IEEE, c2004, 1: 81-88.
[17] LECKIE T, YASINSAC A. Metadata for anomaly-based security protocol attack deduction[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1157-1168.
[18] FADLULLAH Z M, TALEB T, ANSARI N, et al. Combating against attacks on encrypted protocols[C]//In Communications, IEEE International Conference on ICC'07. c2007:1211-1216.
[19] FADLULLAH Z M, TALE B T, VASIAKOS A V, et al. DTRAB: combating against attacks on encrypted protocols through traffic-feature analysis[J]. IEEE/ACM Transactions on Networking (TON), 2010, 18(4): 1234-1247.
[20] YASINSAC A. An environment for security protocol intrusion detection [J]. Journal of Computer Security, 2002, 10(1/2): 177-188.
[21] MAEDCHE A. Ontology learning for the semantic Web[M]. Boston: Kluwer Academic Publishers, 2002.
[22] GAN G, WU J. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm[J]. Pattern Recognition, 2008, 41 (6): 1939-1947.
[23] 朱玉娜, 韩继红, 袁霖, 等. 基于主体行为的多方密码协议会话识别方法[J]. 通信学报, 2015, 11(36): 190-200.
ZHU Y N, HAN J H, YUAN L, et al. Towards session identification using principal behavior for multi-party secure protocol[J]. Journal on Communications, 2015, 11(36): 190-200.
[24] KHAKPOUR A R , LIU A X. High-speed flow nature identification[C]// International Conference on Distributed Computing Systems. Montreal, Canada, c2009: 510-517.

作者简介：

朱玉娜（1985-），女，山东菏泽人，解放军信息工程大学博士生，主要研究方向为安全协议逆向与识别。

韩继红（1966-），女，山西定襄人，博士，解放军信息工程大学教授、博士生导师，主要研究方向为网络与信息安全、安全协议形式化分析与自动化验证。

袁霖（1981-），男，河南商丘人，博士，解放军信息工程大学副教授，主要研究方向为安全协议形式化分析与自动化验证、软件可信性分析。

范钰丹（1982-），女，河南邓州人，解放军信息工程大学讲师，主要研究方向为安全协议形式化分析与自动化验证。

陈韩托（1990-），男，浙江奉化人，解放军信息工程大学硕士生，主要研究方向为协议在线安全性分析。

谷文（1992-），男，湖南圭阳人，解放军信息工程大学硕士生，主要研究方向为安全协议形式化分析与验证。