doi:10.11959/j.issn.1000-436x.2016187

网络加密流量识别研究综述及展望

潘吴斌[1,2]，程光[1,2]，郭晓军[1,2]，黄顺翔[1,2]
(1. 东南大学计算机科学与工程学院，江苏 南京 210096；
2. 东南大学计算机网络和信息集成教育部重点实验室，江苏 南京 210096)

摘 要：鉴于加密流量识别技术的重要性和已有相关研究工作，首先根据流量分析需求的层次介绍了加密流量识别的类型，如协议、应用和服务。其次，概述已有加密流量识别技术，并从多个角度进行分析对比。最后，归纳现有加密流量识别研究存在的不足及影响当前加密流量识别的因素，如隧道技术、流量伪装技术、新型协议 HTTP/2.0 和 QUIC 等，并对加密流量识别趋势及未来研究方向进行展望。

关键词：加密流量识别；网络管理；流量工程；流量伪装；HTTP/2.0
中图分类号：TP393 **文献标识码**：A

Review and perspective on encrypted traffic identification research

PAN Wu-bin[1,2], CHENG Guang[1,2], GUO Xiao-jun[1,2], HUANG Shun-xiang[1,2]
(1. School of Computer Science and Engineering, Southeast University, Nanjing 210096, China;
2. Key Laboratory of Computer Network and Information Integration of Ministry of Education, Southeast University, Nanjing 210096, China)

Abstract: Considering the importance of encrypted traffic identification technology and existing research work, first, the type of encrypted traffic identification according to the demand of traffic analysis were introduced, such as protocols, applications and services. Second, the encrypted traffic identification technology was summarized, and identification technology was compared from multiple views. Third, the deficiencies and the affecting factors of the existing encrypted traffic identification technologies were induced, such as tunneling, traffic camouflage technology, new protocols of HTTP/2.0 and QUIC. Finally, prospect trends and directions of future research on encrypted traffic identification were discussed.

Key words: encrypted traffic identification, network management, traffic engineering, traffic camouflaging, HTTP/2.0

1 引言

流量识别是提升网络管理水平、改善服务质量（QoS）的基础[1]。自“棱镜”监控项目曝光后，全球的加密网络流量不断飙升。Sandvine 报告显示 2015 年 4 月北美的加密流量达到 29.1%，加密互联网访问流量同比上一年增长了 3 倍。当前加密流量快速增长存在多方面原因：1) 用户隐私保护和网络安全意识的增强，安全套接字层（SSL）、安全外壳协议（SSH）、虚拟专用网（VPN）和匿名通信（如 Tor[2]）等技术广泛应用，以满足用户网络安全需求；2) 网络服务提供商（ISP）对 P2P 应用[3]的肆意封堵以及一些公司对即时通信(IM)和流媒体（如 YouTube[4]）等应用的限制，越来越多的应用使用加密和隧道技术应对 DPI（deep packet inspection）技术，以突破这些限制；3) 加密协议良好的兼容性和

收稿日期：2015-12-14；**修回日期**：2016-05-25
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（No.2015AA015603）；江苏省未来网络创新研究院未来网络前瞻性研究基金资助项目（No.BY2013095-5-03）；江苏省“六大人才高峰”高层次人才基金资助项目(No.2011−DZ024)；中央高校基本科研业务费专项资金和江苏省普通高校研究生科研创新计划基金资助项目（No.KYLX15_0118）
Foundation Items: The National High Technology Research and Development Program of China (863 Program) (No.2015AA015603), The Prospective Research Programs Future Internet of Jiangsu Province (No.BY2013095-5-03), The Six Talent Peaks of High Level Talents Project of Jiangsu Province (No.2011−DZ024), The Fundamental Research Funds for the Central Universities and the Research and Innovation Project for College Graduates of Jiangsu Province (No.KYLX15_0118)

可扩展性[5]，采用加密技术变得越来越简单，如现有的 Web 应用可以无缝地迁移到 HTTPS，且 SSL 协议除了能跟 HTTP 搭配，还能跟其他应用层协议搭配（如 FTP、SMTP、POP），以提高这些应用层协议的安全性；4) 采用 HTTPS 加密协议有利于搜索引擎排名，谷歌把是否使用安全加密协议 HTTPS 作为搜索引擎排名的一项参考因素[6]，同等情况下，HTTPS 站点能比 HTTP 站点获得更好的搜索排名。

文献[7~9]综述了当前流量识别的研究进展，虽然取得了不少研究成果，但这些成果大多针对非加密流量识别研究。实际流量识别过程中，加密流量识别与非加密流量识别存在不少差异，主要表现为：1) 由于加密后流量特征发生了较大变化，部分非加密流量识别方法很难适用于加密流量，如 DPI 方法[10]；2) 加密协议常伴随着流量伪装技术（如协议混淆和协议变种[11]），把流量特征变换成常见应用的流量特征；3) 由于加密协议的加密方式和封装格式也存在较大的差异，识别特定的加密协议需要采用针对性的识别方法，或采用多种识别策略集成的方法；4) 当前加密流量识别研究成果主要集中在特定加密应用的识别，实现加密应用精细化识别还存在一定的难度[12]；5) 恶意应用常采用加密技术来隐藏，恶意流量的有效识别事关网络安全。由于缺乏有效的加密流量分析和管理技术，给网络管理与安全带来巨大的挑战，主要表现在以下几个方面。

第一，流量分析和网络管理需要精细化识别加密流量[13]。大多数公司工作时间不允许玩游戏、观看视频和刷微博等娱乐活动。然而，一些员工通过使用加密和隧道技术突破限制。因此，有必要知道加密和隧道协议下运行的具体应用。另外，SSL 协议下运行着各种以 Web 访问为基础的应用，协议下具体运行的应用需要精细化识别(fine-grained identification)，如网页浏览、银行业务、视频或社交网络 SNS。

第二，加密流量实时识别。加密流量识别不仅要识别出具体的应用或服务，还应该具有较好的时效性[14]。比如 P2P 下载和流媒体，实时识别后 ISP 可以提高流媒体的优先级，同时降低 P2P 下载的优先级。

第三，加密通道严重威胁信息安全。恶意软件通过加密和隧道技术绕过防火墙和入侵检测系统[15]将机密信息发送到外网，如僵尸网络[16]、木马和高级持续性威胁（APT）[17]。

2　加密流量识别技术评价指标

目前，加密流量识别主要采用准确性相关指标来进行评估，该指标相对单一，为了满足不断提高的流量分析要求，在原理评估指标基础上，引入兼容性、稳健性、完整性和方向性指标全面、持续地评估加密流量识别技术。下面详细介绍加密流量识别的评估指标。

1) 准确性

准确性反映流量识别技术识别网络应用的能力。

假设 N 为流量样本数，m 为应用类型数，n_{ij} 表示实际类型为 i 的应用被标记为类型 j 的样本数。真正（TP, true positive）代表实际类型为 i 的样本中被正确标记的样本数，$TP_i = n_{ii}$；假正(FP, false positive)代表实际类型为非 i 的样本中被误标识为类型 i 的样本数，$FP_i = \sum_{j \neq i} n_{ji}$。查准率定义为

$$precision = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

假负(FN, false negative)代表实际类型为 i 的样本中被误标识为其他类型的样本数，$FN_i = \sum n_{ij}$。真负（TN, true negative）代表实际类型为非 i 的样本中被标识为非 i 的样本数，$TN_i = n_{jj}$。查全率定义为

$$recall = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

查准率和查全率体现了识别方法在每个单独协议类别上的识别效果。特别是当样本类别分布不均匀时，查全率和查准率可以准确获知每个类别的分类情况。准确率体现了识别方法的总体识别性能，好的算法应该同时具有较高的准确率、查准率和查全率。准确率定义为

$$accuracy = \frac{\sum_{i=1}^{m}\left(TP_i + TN_i\right)}{\sum_{i=1}^{m}\left(TP_i + TN_i + FP_i + FN_i\right)} \tag{3}$$

F-Measure 是综合查准率和查全率得到的评价指标，*F-Measure* 越高表明算法在各个类型的分类性能越好。

$$F\text{-}Measure = \frac{2presion \cdot recall}{presion + recall} \tag{4}$$

2) 完整性

完整性反映了识别方法的识别覆盖率。完整性是指被标识为 i 的样本与实际类型为 i 的样本的比值，相当于查准率和查全率的比值，取值范围可能超过 1。完整性定义为

$$completeness = \frac{recall}{precision} \tag{5}$$

3) 未识别率

未识别率反映流量识别工具对未知流量类型的识别能力。未识别率是指不属于已知流量类型的流量占总流量的比率。F_{total} 表示流量的总字节数或流数，F_{known} 表示被识别流量的字节数或流数。

$$unrecognized = \frac{F_{\text{total}} - F_{\text{known}}}{F_{\text{total}}} \tag{6}$$

4) 兼容性

兼容性反映流量识别工具用于不同网络环境的识别能力。兼容性表示流量识别技术用于不同网络环境的识别能力。acc_j 表示在网络环境 j 的准确率，$\overline{acc}$ 表示平均准确率。

$$portability=\sqrt{\frac{\sum_{j=1}^{m}(acc_j - \overline{acc})^2}{m}} \tag{7}$$

5) 稳健性

稳健性反映流量识别工具长时间维持高识别性能的能力。稳健性表示流量识别技术长时间维持高识别率的能力。acc_k 表示时间段 k 的准确率，acc_0 表示初始准确率。

$$robustness=\sqrt{\frac{\sum_{k=1}^{r}(acc_0 - acc_k)^2}{r}} \tag{8}$$

6) 其他评估指标

另外，一些评估指标进行量化还存在一些问题，如实时性、方向性和计算复杂性。

实时性反映流量识别方法可以在线、快速地识别网络应用的能力。为了及时识别应用，可以根据部分数据分组的特征进行识别，无需等到整条流结束。实时性主要采用流的前若干个分组来体现。

方向性反映流量识别方法对于不同流传输方向的识别能力。IP 流可以分为单向流和双向流，单向流根据传输方向可以分为上行流和下行流，假如第一个数据分组产生分组丢失，无法判断上行和下行方向。方向性可以采用单向流（上行流、下行流）或双向流来体现。

计算复杂性反映流量识别方法准确识别网络应用所需的开销。复杂的识别特征需要耗费大量的存储空间和计算能力，严重影响骨干网的流量分析。计算复杂性可以采用时间和空间复杂度来体现。

3 加密流量识别

广义上来说，加密流量是由加密算法生成的流量。实际上，加密流量主要是指在通信过程中所传送的被加密过的实际明文内容。若用明文 HTTP 协议下载一个加密文件，这种流量不能作为加密流量，因为协议本身是不加密的。近年来，加密流量识别研究取得了一些成果[18]，加密流量识别的首要任务是根据应用需求确定识别对象及识别的类型，再根据识别需求选用合适的识别方法，加密流量识别方法主要可以分为 6 类：基于有效负载检测的分类方法、基于负载随机性的方法、基于数据分组分布的分类方法、基于机器学习的方法、基于主机行为的分类方法和多种策略相结合的混合方法[19]。加密流量识别研究内容框架如图 1 所示。

3.1 识别对象

加密流量识别对象是指识别的输入形式，包括流级、分组级、主机级和会话级，根据流量识别的应用需求确定相应的识别对象，其中，流级和分组级对象使用最广泛，具体描述如下。

流级主要关注流的特征及到达过程，IP 流根据传输方向可以分为单向流和双向流。单向流的分组来自同一方向；双向流包含来自 2 个方向的分组，该连接不一定正常结束，如流超时。有时双向流要求两主机之间从发出 SYN 分组开始到第一个 FIN 分组结束的完整连接。流级特征包括流持续时间、流字节数等。

分组级主要关注数据分组的特征及到达过程，分组级特征主要有分组大小分布、分组到达时间间隔分布等。

主机级主要关注主机间的连接模式，如与主机通信的所有流量，或与主机的某个 IP 和端口通信的所有流量，主机级特征包括连接度、端口数等。

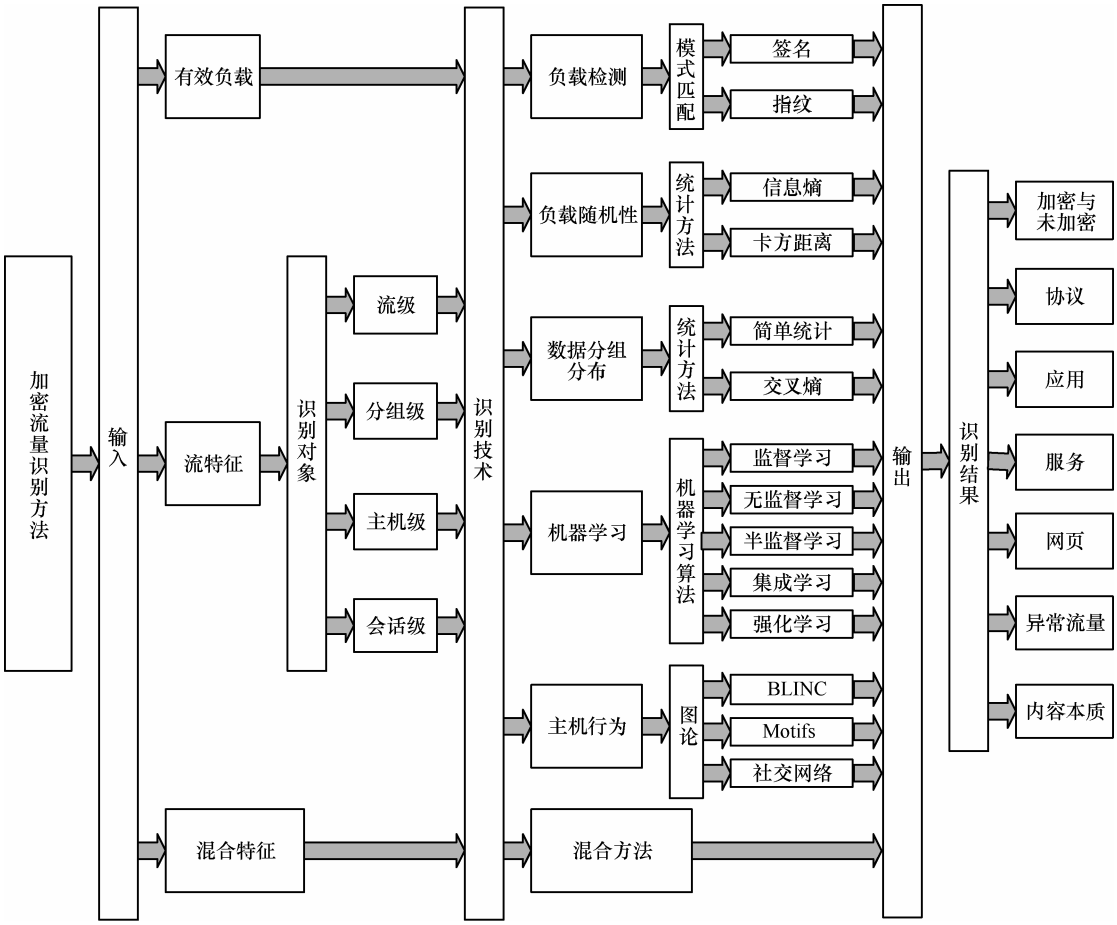


图 1　加密流量识别研究内容

会话级主要关注会话的特征及到达过程，如响应视频请求的数据量较大、针对一个请求会分多个会话传输、会话级特征包括会话字节数和会话持续时间等。

3.2　识别的类型

加密流量识别的类型是指识别结果的输出形式，根据流量识别的应用需求确定识别的类型，加密流量可以从协议、应用、服务等属性逐步精细化识别，最终实现协议识别、应用识别、异常流量识别及内容本质识别等，具体描述如下。

1) 加密与未加密流量，识别出哪些流量属于加密的，剩余则是未加密的。

2) 协议识别就是识别加密流量所采用的加密协议，如 SSL、SSH、IPSec。

3) 应用识别就是识别流量所属的应用程序，如 Skype、BitTorrent 和 YouTube。这些应用还可以进一步精细化分类，如 Skype 可以分为即时消息、语音通话、视频通话和文件传输[20]。

4) 服务识别就是识别加密流量所属的服务类型，如网页浏览（Web browning）、流媒体（streaming media）、即时通信和云存储。

5) 网页识别就是识别 HTTPS 协议下的网页浏览，如百度、淘宝或中国银行等。

6) 异常流量识别就是识别出 DDoS、APT、Botnet 等恶意流量。

7) 内容本质识别就是对应用流量从内容信息上进一步分类，如 YouTube 视频清晰度、音频编码格式，该识别主要用来增加识别透明度。

从以上识别的类型来看，有些流量可能属于一个或多个类型。如 BitTorrent 使用 BitTorrent 协议作为 BitTorrent 网络的应用程序，YouTube 应用程序产生的流量又属于流媒体服务。不同类型的分类表示信息的侧重点不同，因此，流量标签常用于表示不同的类型。下面从加密流量识别的应用需求和识别目标角度详细描述上述识别类型。

3.2.1　加密与未加密流量识别

加密流量识别的首要工作是将加密流量与未加密流量区分，一方面，可以分别采用不同的分类

策略逐步实现精细化识别，另一方面，防止加密应用或服务误识别为非加密流量。另外，一些恶意软件通过加密技术绕过防火墙和入侵检测系统，识别加密流量是异常流量检测的首要任务。文献[21]提出了一种加密流量实时识别方法 RT-ETD，分类器只需要处理每条流中的第一个数据分组，通过第一个数据分组的有效载荷的熵估计就能实时识别加密与未加密流量。实验结果表明该方法识别加密流量的准确率超过 94%，加密流量包括 Skype 和加密 eDonkey，未加密流量的识别准确率高于 99.9%，流量包括 SMTP、HTTP、POP3 和 FTP，因此，该方法可以有效应用于高速网络的加密流量识别。

3.2.2 加密协议识别

加密协议识别由于各协议封装格式不同需要了解协议的交互过程[22]，找出交互过程中的可用于区分不同应用的特征及规律，才有可能总结出网络流量中各应用协议的最佳特征属性[23]，最终为提高总体流识别的粒度与精度奠定基础。加密协议交互过程大体可以分为 2 个阶段，第一阶段是建立安全连接，包括握手、认证和密匙交换，在这过程中通信双方协商支持的加密算法，互相认证并生成密匙；第二阶段采用第一阶段产生的密匙加密传输数据，如图 2 所示。然后，在分析加密协议交互和封装的基础上，详细描述 3 种主流加密协议（IPSec、SSH 和 SSL）的识别。

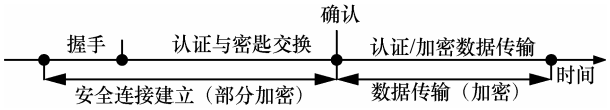


图 2 加密协议一般流程

1) IPSec 安全协议

IPsec 协议是一组确保 IP 层通信安全的协议栈[24]，保护 TCP/IP 通信免遭窃听和篡改以及数据的完整性和机密性。IPSec 协议主要包括网络认证协议 AH（认证头）、ESP（封装安全载荷）和 IKE（密匙交换协议）。AH 协议为 IP 数据分组提供数据完整性保证、数据源身份认证和防重放攻击；ESP 协议包括加密和可选认证的应用方法，除了提供 AH 已有的服务，还提供数据分组和数据流加密以保证数据机密性。IPsec 有隧道和传输 2 种模式，隧道模式中整个 IP 数据分组被用来计算 AH/ESP 头，AH/ESP 头以及 ESP 加密的数据被封装在一个新的 IP 数据分组中，数据封装格式如图 3 所示。

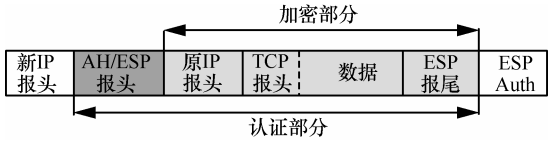


图 3 IPsec 安全协议隧道模式的数据封装格式

由于高实时性应用（如 VoIP 和视频）需要根据识别结果继而提高优先级，使流量识别的实时性变得极为重要。由于隧道和加密技术的应用，基于报头和负载检测的方法无法提供足够的信息来识别应用类型。文献[25]提出一种基于流统计行为的方法可以快速识别 IPSec 隧道下的 VoIP 流量，根据 VoIP 应用的实时性传输需求，介于 60~150 byte 的数据分组较多，实验结果显示该方法可以有效识别 IPSec 隧道中的 VoIP 流量并阻止非 VoIP 流量，从而改善 VoIP 应用的服务质量。

2) SSL/TLS 安全协议

SSL 安全套接层协议提供应用层和传输层之间的数据安全性机制[26]，在客户端和服务器之间建立安全通道，对数据进行加密和隐藏，确保数据在传输过程中不被改变。SSL 协议在应用层协议通信之前就已经完成加密算法和密钥的协商，在此之后所传送的数据都会被加密，从而保证通信的私密性。SSL 协议可以分为 2 层，上层为 SSL 握手协议、SSL 改变密码规则协议和 SSL 警告协议；底层为 SSL 记录协议，SSL 协议分层及数据封装如图 4 所示。

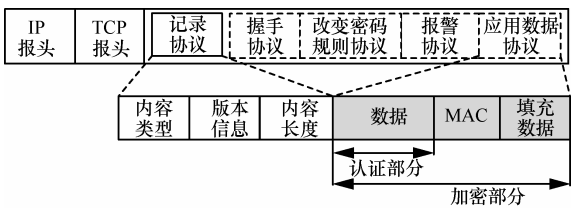


图 4 SSL 安全协议数据封装格式

SSL 加密协议因其良好的易用性和兼容性被广泛应用。文献[27]提出一种针对 SSL 加密应用的识别方法，通过前几个数据分组的大小实现 SSL 流量的早期识别。该方法首先分析 SSL 协议握手期间的报文，根据每个 SSL 记录开始时发送的未加密 SSL 头部的 SSL 配置选项以验证连接是否使用 SSL 协议并确定协议版本，在 SSLv2 协议头部的前 2 位总是 1 和 0，后跟 14 位是 SSL 记录的大小，第 3 个字节是消息类型（1 为“客户端 Hello”，2 为“服务器 Hello”）。SSLv3.0 或 TLS 协议的第 1 个字节

是内容类型（22 为“记录配置”，23 为“有效载荷”），第 2 和第 3 个字节表示主要和次要版本（3 和 0 代表 SSLv3.0，1 代表 TLS），实验结果表明该方法可以达到 85%的识别准确率。

3) SSH 加密协议

SSH 安全外壳协议是一种在不安全网络上提供安全远程登录及其他安全网络服务的协议[28]。SSH 在通信双方之间建立加密通道，保证传输的数据不被窃听，并使用密钥交换算法保证密钥本身的安全。SSH 协议包括传输层协议、用户认证协议和连接协议。传输层协议用于协商和数据处理，提供服务器认证，数据机密性和完整性保护；用户认证协议规定了服务器认证的流程和报文内容；连接协议将加密的安全通道复用成多个逻辑通道，高层应用通过连接协议使用 SSH 的安全机制，SSH 协议分层及数据封装如图 5 所示。

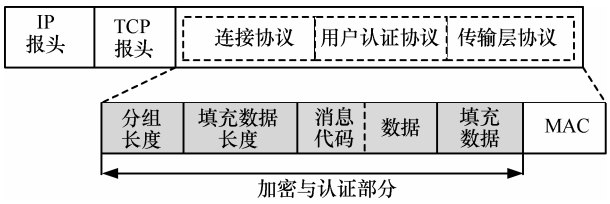


图 5　SSH 安全协议数据封装格式

为了实现不同服务流量(如 VoIP、Video、ERP)的有效管理，快速准确地识别服务类型极其重要。文献[29]提出一种 SSH 实时识别方法，首先通过 SSH 连接的第一个数据分组的统计特征(如到达时间、方向和长度)识别出 SSH 协议流量，然后通过 *k*-means 聚类分析前 4 个数据分组的统计特征识别 SSH 协议承载的应用(如 SCP、SFTP 和 HTTP)。实验结果表明，该方法 SSH 流量的识别准确率达到 99.2%，同时，SSH 协议下应用的识别准确率达到 99.8%。

3.2.3　服务识别

服务识别就是识别加密流量所属的应用类型，属于 Video Streaming 的应用有 YouTube、Netflix、Hulu 和 Youku 等，属于 SNS 的应用有 Facebook、Twitter 和 Weibo 等，属于 P2P 的应用有 Skype、BitTorrent 和 PPlive 等。此外，具体应用还可以进一步精细化识别，如 Skype 可以分为即时消息、语音通话、视频通话和文件传输。

随着 P2P 应用的不断发展，P2P 流量占据了当今互联网流量较大的份额，而且 P2P 应用大都使用混淆技术，如动态端口号、端口跳变、HTTP 伪装和负载加密，因此，为了实现更好的 P2P 流量管控，需要有效的 P2P 流量识别方法。文献[30]比较了 3 种 P2P 流量识别方法，包括基于端口、应用层签名和流统计特征识别方法。实验结果表明基于端口的方法无法识别 30%~70%的流量，应用层签名方法准确性较高，但可能因法律或技术原因无法使用，而流统计特征识别方法可以达到 70%的识别率。BitTorrent 是常用的 P2P 应用，占 P2P 流量相当大的比例，文献[31]使用统计分析方法识别 BitTorrent 流量与其他类型的流量(如类似功能的 FTP)，实验结果表明该方法可以有效地实时识别 BitTorrent 流量。Skype 由于出色的声音质量和易用性被看作最好的 VoIP 软件，引起了研究界和网络运营商的广泛关注。由于 Skype 应用封闭源代码无法获知协议和算法，且强大的加密机制使识别难度较大。文献[32]提出了一种名为 Skype-hunter 的方法实时识别 Skype 流量，该方法通过基于签名和流统计特征相结合的策略能够有效识别信令业务以及数据业务（如语音、视频和文件传输）。

当前，P2P-TV 应用无论用户数还是所产生的流量都是互联网上增长最快的，实现在互联网上以较低的费用实时观看电视节目。由于大部分 P2P-TV 应用是基于专有或未知的协议，使 P2P-TV 流量识别难度大。文献[33, 34]提出一种识别 P2P-TV 应用的新方法，该方法根据较小的时间窗与其他主机交换的数据分组数和字节特征，这 2 种特征包含了应用和内部运作多方面的信息，如信令行为和视频块大小，该方法采用支持向量机算法可以准确地识别 P2P-TV 应用，再通过简单计算数据分组数量还可以精细化识别 P2P-TV 应用（如 PPLive、SopCast、TVAnts 和 Joost），最终，数据分组和字节准确率均高于 81%。

3.2.4　异常流量识别

虽然研究已经表明基于信息熵的方法在异常检测中可以取得较好的性能，但还未有研究采用熵检测与流量分布相结合的方法。文献[35]提出一种根据流头部特征（IP 地址、端口和流字节数）和行为特征（测量与每个主机通信的目的/源 IP 数量的度分布）分布相结合的方法，发现地址和端口分布的熵值相关性强，异常检测能力非常相似；行为和流字节数分布不太相关，可以检测出根据端口和地址分布无法检测的异常流量，实验结果表明端口和

地址分布在检测扫描和泛洪异常能力有限，而基于熵的异常检测方法具有较好的识别性能。Lakhina 等[36]采用数据分组的特征（IP 地址和端口）分布对大范围异常流量进行检测和识别，基于熵的识别方法可以非常灵敏地检测到大范围异常流量，还可以通过无监督学习自动识别异常流量。实验结果表明聚类方法可以有效地将正常流量和异常流量分为不同的集群，可以用来发现新的异常流量。Soule 等[37]提出一种采用基于流量矩阵的方法识别异常流量，首先采用卡尔曼过滤器识别出正常流量，然后采用 4 种不同方法（阈值、方差分析、小波变换和广义似然比）识别异常流量。实验结果的 ROC 曲线表明这些方法可以较好地实现误报和漏报的平衡。

3.2.5 内容本质识别

内容本质识别就是对加密应用流量从内容本质上进一步识别，如视频清晰度、图片格式。文献[38]首次提出一种内容本质的识别方法 Iustitia，基本思路是统计特定数量的连续字节的熵，再采用机器学习方法进行识别。首先，基于文本流熵值最低，加密流熵值最高，以及二进制流熵值处于中间的特性，实现文本流，二进制流和加密流的实时识别。然后，进一步扩展 Iustitia 方法实现二进制流的精细化识别，区分不同类型的二进制流（如图像、视频和可执行文件），甚至可以识别二进制流传输的文件类型（如 JPEG 和 GIF 图像、MPEG 和 AVI 视频）。实验使用 1 KB 大小缓存的识别精度可以达到 88.27%，且 91.2%的流的识别时间不超过 10%的分组到达时间间隔，表明该方法具有较高的识别效果和效率。

3.3 加密流量识别方法

虽然传统流量识别研究取得了不少成果[39, 40]，但有些识别方法很难适用于加密流量，加密流量识别的前提是针对不同的应用或协议有明显的区分特征。加密流量识别和未加密流量识别的本质区别在于由于加密使用于区分的特征发生了改变，流量加密后的变化可以概括如下：1) IP 报文的明文内容变为密文；2) 流量加密后负载的统计特性（如随机性或熵）发生改变；3) 流量加密后流统计特征发生改变，如流字节数、分组长度和分组到达时间间隔。这些变化使有些传统识别方法很难或根本无法适用，如基于端口和应用层负载的方法，下面详细介绍当前主流的加密流量识别技术，并从成本代价、识别速度、识别粒度等方面进行对比分析。

3.3.1 基于有效负载的识别方法

基于有效负载的识别方法[41]通过分析数据分组的有效负载来识别流量，但该方法由于解析数据分组负载触犯隐私，处理私有协议或加密协议时由于应用层数据加密可能很难起作用，且当协议发生变化时必须同步更新。然而，一些加密协议在密匙协商过程中数据流是不加密或部分加密的，可以从这部分未加密数据流中提取有用的信息来识别协议或应用。文献[42]提出一种 Skype 流量实时识别的框架，由于数据分组的协议报头未加密，可以通过统计协议报头前 4 个字节的卡方确定数据流的具体协议，甚至可以根据 Skype 流量的随机特征（分组到达速率和分组长度），采用贝叶斯分类器识别出该协议所承载的应用类型（文件传输、音频和视频等）。Korczynski[43]提出一种基于马尔可夫链的随机指纹方法，该方法对 SSL/ TLS 会话从服务器到客户端的 SSL / TLS 协议头部的消息序列建立一阶马尔可夫链，马尔可夫链就是该会话的指纹。实验结果表明该方法具有较好的识别性能，同时发现许多应用未严格遵循 RFC 规范，识别时需要根据协议更新等变化适当调整指纹。

3.3.2 数据分组负载随机性检测

负载随机性检测方法根据网络应用的数据流并不完全随机加密的特性进行识别，由于每个数据分组会携带一些相同的特征字段，所以数据分组的这些字节可能不是随机的，可以根据这些特征字段的随机性来识别。文献[44]提出一种基于加权累积和检验的加密流量盲识别方法，该方法利用加密流量的随机性，对负载进行累积和检验，根据报文长度加权综合，最终实现在线普适识别，实验结果显示加密流量识别率达到 90%以上。文献[38]提出的 Iustitia 识别方法根据文本流熵值最低，加密流熵值最高，以及二进制流熵值介于两者之间的特性，采用基于熵值的方法精细化识别二进制流（如图像、视频和可执行文件），甚至可以识别二进制流传输的文件类型（如 JPEG 和 GIF 图像、MPEG 和 AVI 视频）。

3.3.3 基于机器学习的识别方法

加密技术只对载荷信息进行加密而不对流统计特征进行处理，因此，基于流统计特征的机器学习识别方法[45]受加密影响较小。Okada[46]在分析流量加密导致特征变化的基础上，提出一种基于特征估计的识别方法 EFM，根据未加密与加密流量的相

关性选取强相关特征，实验结果表明该方法可以达到 97.2%的识别准确率。Alshammari[47]采用多种监督学习方法(如 C4.5、AdaBoost、GP、SVM、RIPPER、Naïve Bayes)用于加密与非加密 SSH 和 Skype 流量的识别，实验结果表明机器学习方法具有较好的识别性能，且可以较好地适用于不同网络环境。Korczynski[48]提出一种统计方法识别 Skype 流量的服务类型（如语音通话、SkypeOut、视频会议、聊天）。使用前向特征选择方法选取最优特征子集，尽管语音和视频流量识别相对较难，但该方法仍取得较高的识别精度。Erman[49]首次将半监督学习方法用于网络流量识别，分类器通过不断的迭代学习可以识别未知的和行为有变化的应用。文献[50]采用子空间聚类方法针对不同分类器使用独立的特征子集单独识别每种应用，而不是采用统一的特征子集来识别应用。文献[51]提出一种针对 Tor 应用的识别方法，该方法首先选取应用的代表性流特征，如爆发量和方向，再采用隐马尔可夫模型识别 Tor 应用承载的流量。

3.3.4　基于行为的识别方法

首先，基于行为的识别方法[52]是从主机的角度来分析不同应用的行为特征，识别结果通常是粗粒度的，如 P2P 和 Web。其次，该方法对于传输层加密无能为力。第三，使用网络地址转换（NAT）和非对称路由等技术会因为不完整的连接信息而影响其识别性能。基于行为的识别方法可以分为主机行为和应用行为。基于主机行为的方法针对加密协议更新和新协议顽健性高，可以用于骨干网实时粗粒度识别[53, 54]。Karagiannis[52]提出了一种基于主机行为的流量识别方法 BLINC，该方法是一种识别未知流量的启发式技术，尝试获得主机的参数，一旦这些主机的连接被建立，连接到已知主机上的流量可以被简单地标注为已知主机正在使用的应用。Schatzmann[55]利用主机和协议的相关性，以及周期行为特点从 HTTPS 流量中识别加密 Web 邮件。文献[56]提出一种在时间窗口内根据交互数据分组的数目和字节数实现 P2P-TV 流量精细化识别。Xiong[57]提出一种基于主机行为关联的加密 P2P 流量实时识别方法。基于某些先验知识，节点和节点之间的连接、节点和服务器之间的连接等通信模式来识别 P2P 流量。虽然基于应用行为的方法根据应用周期性的操作和通信模式能有效地实现精细化识别，但实际上只有部分加密应用可以适用。

3.3.5　基于数据分组大小分布的识别方法

在实际网络环境中，为提高用户体验，服务提供商会针对不同的业务类型对数据流中的数据分组大小进行处理，如流媒体的数据分组不宜过大，否则网络拥塞时影响播放流畅度，而文件下载的数据分组通常以最大负载传输。因此，可以根据业务类型数据分组大小分布差异进行识别，该方法受加密影响较小。文献[58]提出一种基于数据分组大小分布签名的新方法，该方法首先根据双向流模型将流量分组聚集成双向流，从而获取不同终端之间的交互行为特征，再使用分组大小分布的签名获取双向流中分组的负载大小分布（PSD）的概率，最后，采用 Renyi 交叉熵计算双向流和应用的 PSD 之间的相似性来进行识别。该方法在减少数据分组处理量的同时实现 P2P 和 VoIP 应用的准确识别。

3.3.6　混合方法

由于很多识别方法只对特定协议有效，因此可以将多种加密流量识别方法集成实现高效的加密流量识别。文献[59]提出一种签名和统计分析相结合的加密流量识别方法，首先采用特征匹配方法识别 SSL/TLS 流量，然后应用统计分析确定具体的应用协议，实验结果表明该方法能够识别 99%以上的 SSL/TLS 流量，*F-Measure* 达到 94.52%。文献[60]提出了一种 P2P 流量的细粒度识别方法，该方法通过统计特定 P2P 应用中经常且稳定出现的特殊流。该方法利用流的几个通用特性就可以达到较高的识别准确率；其次，即使待识别的应用混杂其他高带宽消耗的应用也可以很好地识别，性能表现优于大多数现有的主机识别方法。文献[61]提出一种结合多个流量识别算法的混合方法，通过 4 种不同的组合机制在 4 个不同的网络场景下进行验证，实验结果表明混合方法在不同场景下都具有较好的识别率和稳健性。

3.3.7　加密流量识别方法综合对比

在上述加密流量识别方法中，很难借助一种方法识别所有的流量，大部分方法适用于特定的协议或应用，因此，有必要对上述加密流量识别方法从不同角度进行深入分析和对比，如表 1 所示。表 2 概述了加密流量识别研究的具体实例，包括识别对象、识别特征和识别方法等。文献[62]比较了分组头特征(如签名、指纹)和流统计特征所获得的识别性能，结果表明选择性集成 2 组特征可以获得更快

表 1 加密流量识别方法对比

方法	检测内容	成本代价	识别速度	实时性	识别粒度	准确性	未识别率	兼容性	稳健性
负载随机性	部分负载	★★★★★	★★☆☆☆	★☆☆☆☆	★★★☆☆	★★☆☆☆	★★☆☆☆	★★★★☆	★★☆☆☆
有效负载	负载	★★★★★	★★☆☆☆	★☆☆☆☆	★★★☆☆	★★★★★	★★☆☆☆	★★★★☆	★★★☆☆
机器学习	流统计特征	★★☆☆☆	★★★★★	★★★★☆	★★★★★	★★★★☆	★★★★☆	★★☆☆☆	★★☆☆☆
行为	主机行为	★★★☆☆	★★★★☆	★★★☆☆	★★★★☆	★★★☆☆	★★★☆☆	★★☆☆☆	★★★☆☆
数据分组大小	数据分组大小	★★☆☆☆	★★★★★	★★★★☆	★★★★☆	★★★★☆	★★★☆☆	★★★☆☆	★★★☆☆
混合方法	多种特征	★★★☆☆	★★★☆☆	★★★☆☆	★★★★★	★★★★★	★★★★☆	★★★★☆	★★★★☆

表 2 加密流量识别研究实例概述

实例	识别对象	特征	识别方法	算法	数据集	标记
文献[27]	SSL & non-SSL SSL 协议下应用	分组大小	机器学习	基于 GMM 聚类	P6-2004,2006, UMass	已知
文献[38]	加密与未加密,不同加密算法	字符随机性	负载随机性	熵矩阵估计	Campus	签名
文献[42]	Skype 及 Skype 协议下应用	签名	有效负载	签名、机器学习	Campus, ISP	已知
文献[43]	SSL 协议下应用	指纹	混合方法	指纹、HMM	私有	签名
文献[47]	SSH & non-SSH、Skype & non-Skype	分组头特征 流特征	机器学习	C4.5、AdaBoost、GP	MAWI, DARPA99	端口, Packet Shaper
文献[52]	Edonkey、MSN、SSH 等	行为特征	主机行为	启发式算法	GN, UN1,2	签名
文献[55]	HTTPS、Tor、Oscar 等	签名、流特征	混合方法	匹配算法、NB	DARPA,私有	已知
文献[58]	P2P、VoIP	数据分组大小	分组大小分布	Reiyi 交叉熵	CERNET	手动标记
文献[64]	SSH、SSL 及非加密应用	流特征	机器学习	改进的 *k*-means	Campus,公共	L7-filter,端口
文献[65]	SSH 协议下应用	分组大小、方向	机器学习	GMM、SVM	私有	SSHgate
文献[66]	BitTorrent 与非加密应用	流特征	机器学习	*k*-means 和 KNN 混合	私有	Cisco SCE 2020 box
文献[67]	SSH、HTTPS 及非加密应用	分组大小、到达时间间隔、方向	机器学习	Profile HMM	GMU	端口
文献[68]	SSH 及非加密应用	行为特征	主机行为	图论	LBL, GMU	端口
文献[69]	Skype	指纹、流特征	混合方法	Chi-Square、NB	Campus, ISP	已知

和更准确的识别性能，使用所有可用的特征并不会获得最好的识别性能[63]。

4 挑战及未来研究方向

4.1 主要问题

从前述内容可知，加密流量识别技术已引起国内外研究人员的极大关注，已成为网络管理领域一项重要的研究内容。尽管当前加密流量识别通过机器学习等方法取得了不少研究成果，但仍然存在潜在问题尚未很好解决，加密流量识别还存在以下几个方面的问题。1) 算法评估及比较，评估需要当前最新的数据集，不同算法的比较又需要相同的数据集，由于隐私和知识产权等原因很难公开或共享算法和数据集。2) 流量分类器的可扩展能力，由于网络流量爆发式增长，当前研究大都基于小规模流量的测试，应用于真实环境还有一定的距离，特别是骨干网。3) 样本标记，当前大多数样本标记工具都是基于 DPI，该技术标记加密流量的能力有限；另外，公开数据集很少提供标记信息，且不带负载。4) 流量分类器的兼容性和稳健性。分类器需要在不同的网络环境进行长时间的验证，分类器可能在某个环境具有较好的识别性能；另外，随着时间推移，应用协议在不断变化，分类器也需要不断的更新。5) 流量伪装问题。越来越多的加密技术采用流量伪装，把流量伪装成其他常用应用的特征防止识别，如匿名攻击和 P2P 下载。

4.2 加密流量识别的影响因素

1) 隧道技术

隧道技术常用于在不安全的网络上建立安全通道，将不同协议的数据分组封装然后通过隧道发送，识别隧道协议相对容易，识别隧道协议下承载的应用相对较难，因此，加密流量识别要考虑隧道技术的影响。隧道技术的数据分组格式主要由传输

协议、封装协议和乘客协议组成。乘客协议就是用户数据分组必须遵守内网的协议，而隧道协议则用于封装乘客协议，负载隧道的建立、保持和断开。常见的隧道协议有基于 IPSec 的第二层隧道协议(L2TP)或基于 SSL 的安全套接字隧道协议(SSTP)。L2TP 依靠 IPSec 协议的传输模式来提供加密服务，L2TP 和 IPSec 的组合称为 L2TP/IPSec。SSTP 协议提供了一种用于封装 SSL 通道传输的 PPP 通信机制，SSL 提供密钥协商、加密和完整性检查，确保传输安全性。

2) 代理技术

加密流量识别还要克服代理技术带来的影响，数据压缩代理技术可以有效减少带宽使用，但对流量识别产生较大的影响。如 Chrome 浏览器采用数据压缩代理技术，能够有效提高网页加载速度，并节省网络流量。当用户使用代理功能时，Google 服务器会对 Web 请求的内容进行压缩和优化处理，由于 Chrome 浏览器与服务器之间采用 SPDY 协议，该协议会对内容进一步优化处理，使用数据压缩代理技术可以节约用户 50%的数据流量。然而，数据压缩使流统计特征发生较大的改变，使基于流统计特征的方法识别性能下降，如何维持识别方法的稳健性还有待进一步研究。

3) 流量伪装技术

流量伪装技术（如协议混淆、流量变种）将一种流量的特征伪装成另一种流量（如 HTTP），降低基于流统计特征方法的识别准确率，且木马蠕虫等恶意程序也越来越多的采用流量伪装技术进行恶意攻击和隐蔽通信。Wright[70]提出了一种实时修改数据分组的凸优化方法，将一种流量的分组大小分布伪装成另一种流量的分组大小分布，变换后的流量可以有效地躲避 VoIP 和 Web 等流量分类器的识别。另外，还有一种称为匿名通信的流量伪装技术，用于隐藏网络通信中发送方与接收方的身份信息（如 IP 地址）以及双方通信关系，通过多次转发和改变报文的样式消除报文间的对应关系，从而为网络用户提供隐私保护。Tor 是目前应用最广泛的匿名通信系统，使用 Tor 系统时，客户端会选择一系列的节点建立 Tor 链路，链路中的节点只知道其前继节点和后继节点，不知道链路中的其他节点信息。

4) HTTP/2.0、SPDY 及 QUIC 协议

为了降低延迟和提高安全性，Google 推出 SPDY 协议替代 HTTP 协议。SPDY 协议采用多路复用技术，可在一个 TCP 连接上传送多个资源，且优先级高的资源优先传送，并强制采用 TLS 协议提高安全性。Google 为了支持 IETF 提出的 HTTP/2.0 协议成为标准弃用 SPDY 协议，HTTP/2.0 协议实际上是以 SPDY 协议为基础，采用 SPDY 类似的技术，如多路复用技术和 TLS 加密技术。由于 HTTP/2.0 协议基于 TCP 仍然存在时延问题，Google 提出一种基于 UDP 的传输层协议 QUIC（quick UDP internet connection）。QUIC 协议结合 TCP 和 UDP 协议的优势，解决基于 TCP 的 SPDY 协议存在的瓶颈，实现低延时、高可靠性和安全性。

4.3 未来研究方向

综上讨论，有些问题是流量识别的共同问题，不只出现在加密流量识别研究中。今后可从以下几个方面展开科研工作。

1) 加密流量精细化识别

随着流量分析需求的提高，为了加强流量管控，识别流量是否加密是远远不够的，因为实际网络管理中需要识别加密协议或隧道协议下的应用或服务。要实现精细化识别这一目标，多阶段逐步精细化识别和混合方法是较好的解决思路，在各个阶段中完成不同的识别任务，或结合不同的算法识别不同的应用[71]。

2) SSL 协议下的应用识别

由于 SSL 协议较好的兼容性和易用性，越来越多的网络应用都积极使用 SSL 协议来确保通信过程的安全性。SSL 协议在网页浏览、观看视频和社交网络等广泛应用，使该流量呈爆发式增长，且基于 SSL 协议的应用变得越来越复杂。SSL 协议在保护用户隐私和数据安全的同时也给网络管理提出了更高的要求，如何精细化识别 SSL 协议下的网络应用已成为当前网络管理面临的挑战。

3) 加密视频内容信息识别

随着视频业务应用越来越广泛，视频流量占比不断增加，网络运营商和视频服务提供商需要知道当前的视频体验业务质量，从而改善视频 QoS。YouTube 作为最常用的视频网络，90%以上流量采用加密技术，国内视频网络也越来越多的采用加密技术。在加密场景下，很难获取与视频体验业务质量相关的参数（如播放码率和清晰度）[72]。因此，如何识别加密视频码率和清晰度对评估和改善 QoS 具有重要意义。

4) 加密流量数据集的准确标记

近年来，一些新的算法和技术已被提出，具有较好的分类性能。然而，由于收集的网络流量不同分类结果无法直接比较。无论是建立机器学习分类模型，还是验证分类模型性能都需要标记数据集。公共数据集大都没有有效载荷信息和标记信息，加密流量哪怕有效载荷也很难通过 DPI 工具（如 nDPI[73]、Libprotoident[74]）进行标记。因此，一些研究人员不得不借助常用端口号再增加过滤规则进行标记，导致基准不准确[75]。此外，为了满足加密流量的精细化识别要求，关键是要标记加密协议下运行的不同应用，使标记难上加难。自生成数据集主要采用监视主机内核的方式[76]或 DPI 方式获取标记，自生成数据集虽然相对容易获取标记信息，但各自采用自生成数据集会造成不同算法间无法对比的问题。

5) 流量伪装

基于流特征的识别方法是加密流量识别最常用的方法，因此，也在不停地研究相应的流量模式伪装技术，如流量填充、流量规范化和流量掩饰，通过伪装技术将一种流量的特征伪装成另一种流量[77]，模糊流量特征以降低流量识别性能。Wright[70]提出了一种实时修改数据分组的凸优化方法，将一种流量的分组大小分布伪装成另一种流量的分组大小分布，变换后的流量可以有效地躲避 VoIP 和 Web 等流量分类器的识别。未来流量伪装技术将集成流量填充、流量规范化和流量掩饰等多种手段应对流量分析，且流量伪装的多样性和自适应能力将大大增强。除此之外，匿名通信、隧道技术和代理技术都是流量伪装的不同表现形式，匿名通信通过隐藏网络通信中发送方与接收方的身份信息以及双方通信关系防止追踪，隧道技术通过 L2TP、SSTP 等协议对数据分组再封装，而数据压缩代理技术为了节约流量使流统计特征发生变化。因此，需要改进目前的识别方法应对即将到来的挑战。

6) 协议出新及业务分布变化

由于应用协议的改进和优化，以及为阻碍流量识别会不断推出新版本，随之协议签名和行为特征发生改变，因此，原有的识别方法需要周期性更新。随着用户对网络安全和网络性能需求的提高，SPDY、HTTP/2.0 及 QUIC 等新加密协议不断推出，用于解决基于 TCP 及 UDP 的协议存在的瓶颈，实现低延时、高可靠和安全的网络通信。在不久的将来，HTTP/2.0 和 QUIC 协议将被广泛应用，如何识别协议下承载的应用面临新的挑战。

另外，机器学习分类方法因其不受加密影响广泛应用于加密流量识别，但基于机器学习的识别方法会因为不同时间段以及不同地域的流量所承载的业务分布差异而引起概念漂移问题[78, 79]。大多数算法只在一个或 2 个特定场景下表现良好，不同算法在不同场景的识别能力差异较大。根据不同场景的流量训练的分类器对新样本的适用性逐渐变弱，导致分类模型的识别能力下降。如果能够准确地识别流量概念漂移，就可以及时有效地更新分类器，从而避免频繁更新分类器。

5 结束语

加密流量识别是当前流量识别领域最具挑战性的问题之一。本文首先介绍加密流量识别的研究背景及意义。其次，阐述加密流量识别对象及识别的类型，加密流量识别根据应用需求从加密与未加密识别逐渐精细化识别的过程，包括协议识别、应用类型识别、内容本质识别和异常流量识别。然后，综述当前加密流量识别方法并对比分析，可以看出多阶段或多方法集成的混合方法是未来的研究热点。接着，阐述当前加密流量识别的影响因素，包括隧道技术、代理技术、流量伪装和新协议 HTTP/2.0 与 QUIC。最后，归纳现有加密流量识别的不足，展望加密流量识别趋势和未来的研究方向，可以从以下几方面开展：1) 采用 2 种或多种方法集成的多层分级框架进行加密流量精细化识别；2) 实现 SSL 协议下应用的精细化识别，以及 HTTPS 加密视频流量的内容识别(如视频码率和清晰度)；3) 建立大规模具备细粒度标记的加密流量数据集；4) 研究更有效的流量识别技术应对流量伪装等反措施；5) 建立自适应应用协议出新及流量分布变化的分类模型及技术。

参考文献：

[1] ROUGHAN M, SEN S, SPATSCHECK O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[C]//The 4th ACM SIGCOMM Conference on Internet measurement. ACM, 2004: 135-148.

[2] DINGLEDINE R, MATHEWSON N, SYVERSON P. Tor: the second-generation onion router[R]. Naval Research Lab Washington DC, 2004.

[3] GOMES J V, INÁCIO P R M, PEREIRA M, et al. Detection and classification of peer-to-peer traffic: a survey[J]. ACM Computing Sur-

veys (CSUR), 2013, 45(3): 30.

[4] GILL P, ARLITT M, LI Z, et al. Youtube traffic characterization: a view from the edge[C]//The 7th ACM SIGCOMM Conference on Internet Measurement. ACM, 2007: 15-28.

[5] ZHANG X B, LAM S S, LEE D Y, et al. Protocol design for scalable and reliable group rekeying[J]. IEEE/ACM Transactions on Networking, 2003, 11(6): 908-922.

[6] BARRY S. Google starts giving a ranking boost to secure HTTPS/SSL sites [EB/OL]. http://searchengineland.com/google-starts-giving- ranking-boost-secure-httpsssl-sites-199446, 2015.

[7] NGUYEN T T T, ARMITAGE G. A survey of techniques for internet traffic classification using machine learning[J]. Communications Surveys & Tutorials, IEEE, 2008, 10(4): 56-76.

[8] NAMDEV N, AGRAWAL S, SILKARI S. Recent advancement in machine learning based internet traffic classification[J]. Procedia Computer Science, 2015, 60: 784-791.

[9] DAINOTTI A, PESCAPE A, CLAFFY K C. Issues and future directions in traffic classification[J]. Network, IEEE, 2012, 26(1): 35-40.

[10] BUJLOW T, CARELA-ESPAÑOL V, BARLET-ROS P. Independent comparison of popular DPI tools for traffic classification[J]. Computer Networks, 2015, 76: 75-89.

[11] WRIGHT C V, COULL S E, MONROSE F. Traffic morphing: an efficient defense against statistical traffic analysis[C]//NDSS. 2009: 237-250.

[12] VELAN P, ČERMÁK M, ČELEDA P, et al. A survey of methods for encrypted traffic classification and analysis[J]. International Journal of Network Management, 2015, 25(5): 355-374.

[13] PARK B, HONG J W K, WON Y J. Toward fine-grained traffic classification[J]. Communications Magazine, IEEE, 2011, 49(7): 104-111.

[14] BERNAILLE L, TEIXEIRA R, AKODKENOU I, et al. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26.

[15] FADLULLAH Z M, TALEB T, VASILAKOS 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.

[16] GU G, ZHANG J, LEE W. BotSniffer: detecting botnet command and control channels in network traffic[C]//Network and Distributed System Security Symposium. 2008.

[17] TANKARD C. Advanced persistent threats and how to monitor and deter them[J]. Network Security, 2011, 2011(8): 16-19.

[18] CAO Z, XIONG G, ZHAO Y, et al. A survey on encrypted traffic classification[M]//Applications and Techniques in Information Security. Springer Berlin Heidelberg, 2014: 73-81.

[19] GRIMAUDO L, MELLIA M, BARALIS E. Hierarchical learning for fine grained internet traffic classification[C]//Wireless Communications and Mobile Computing Conference (IWCMC). IEEE, 2012: 463-468.

[20] ROSSI D, VALENTI S. Fine-grained traffic classification with netflow data[C]//The 6th International Wireless Communications and Mobile Computing Conference. ACM, 2010: 479-483.

[21] DORFINGER P, PANHOLZER G, JOHN W. Entropy estimation for real-time encrypted traffic identification (short paper)[M]. Springer Berlin Heidelberg, 2011.

[22] BELLOVIN S M, MERRITT M. Cryptographic protocol for secure communications: U.S. Patent 5,241,599[P]. 1993-8-31.

[23] FAHAD A, TARI Z, KHALIL I, et al. Toward an efficient and scalable feature selection approach for internet traffic classification[J]. Computer Networks, 2013, 57(9): 2040-2057.

[24] Kent security architecture for the internet protocol[EB/OL]. https://tools.ietf.org/html/rfc4301, 2015.

[25] YILDIRIM T, RADCLIFFE P J. VoIP traffic classification in IPSec tunnels[C]//Electronics and Information Engineering (ICEIE). IEEE, 2010, 1: V1-151-V1-157.

[26] DIERKS T. The transport layer security (TLS) protocol version 1.2 [EB/OL]. https://tools.ietf.org/html/rfc5246, 2015.

[27] BERNAILLE L, TEIXEIRA R. Early recognition of encrypted applications[M]//Passive and Active Network Measurement. Springer Berlin Heidelberg, 2007: 165-175.

[28] YLONEN T. The secure shell (SSH) transport layer protocol[EB/OL]. https://tools.ietf.org/html/rfc4253, 2015.

[29] MAIOLINI G, BAIOCCHI A, IACOVAZZI A, et al. Real time identification of SSH encrypted application flows by using cluster analysis techniques[C]//NETWORKING 2009. Springer Berlin Heidelberg, 2009: 182-194.

[30] MADHUKAR A, WILLIAMSON C. A longitudinal study of P2P traffic classification[C]//Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS 2006. IEEE, 2006: 179-188.

[31] LE T M, BUT J. Bittorrent traffic classification[R]. Centre for Advanced Internet Architectures. Technical Report A, 91022.

[32] ADAMI D, CALLEGARI C, GIORDANO S, et al. Skype‐hunter: a real‐time system for the detection and classification of Skype traffic[J]. International Journal of Communication Systems, 2012, 25(3): 386-403.

[33] VALENTI S, ROSSI D, MEO M, et al. Accurate, fine-grained classification of P2P-TV applications by simply counting packets[M]//Traffic Monitoring and Analysis. Springer Berlin Heidelberg, 2009: 84-92.

[34] BERMOLEN P, MELLIA M, MEO M, et al. Abacus: accurate behavioral classification of P2P-TV traffic[J]. Computer Networks, 2011, 55(6): 1394-1411.

[35] NYCHIS G, SEKAR V, ANDERSEN D G, et al. An empirical evaluation of entropy-based traffic anomaly detection[C]//The 8th ACM SIGCOMM Conference on Internet Measurement. ACM, 2008: 151-156.

[36] LAKHINA A, CROVELLA M, DIOT C. Mining anomalies using traffic feature distributions[J]. ACM SIGCOMM Computer Communication Review, 2005, 35(4): 217-228.

[37] SOULE A, SALAMATIAN K, TAFT N. Combining filtering and statistical methods for anomaly detection[C]//The 5th ACM SIGCOMM Conference on Internet Measurement. USENIX Association, 2005: 31.

[38] KHAKPOUR A R, LIU A X. An information-theoretical approach to high-speed flow nature identification[J]. IEEE/ACM Transactions on Networking (TON), 2013, 21(4): 1076-1089.

[39] CALLADO A, KAMIENSKI C, SZABÓ G, et al. A survey on internet traffic identification[J]. Communications Surveys & Tutorials, IEEE, 2009, 11(3): 37-52.

[40] KIM H, CLAFFY K C, FOMENKOV M, et al. Internet traffic classification demystified: myths, caveats, and the best practices[C]// Proceedings of the 2008 ACM CoNEXT Conference. ACM, 2008: 11.

[41] FINSTERBUSCH M, RICHTER C, ROCHA E, et al. A survey of payload-based traffic classification approaches[J]. Communications Surveys & Tutorials, IEEE, 2014, 16(2): 1135-1156.

[42] BONFIGLIO D, MELLIA M, MEO M, et al. Revealing skype traffic: when randomness plays with you[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(4): 37-48.

[43] KORCZYNSKI M, DUDA A. Markov chain fingerprinting to classify encrypted traffic[C]//INFOCOM, 2014 Proceedings IEEE. IEEE, 2014: 781-789.

[44] 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.
ZHAO B, GUO H, LIU Q R, et al. Protocol independent identification of encrypted traffic based on weighted eumnlative sum test[J]. Journal of Software, 2013, 24(6): 1334-1345.

[45] MOORE A W, ZUEV D. Internet traffic classification using bayesian analysis techniques[J]. ACM SIGMETRICS Performance Evaluation Review. 2005, 33(1): 50-60.

[46] OKADA Y, ATA S, NAKAMURA N, et al. Comparisons of machine learning algorithms for application identification of encrypted traffic[C]//Machine Learning and Applications and Workshops (ICMLA). IEEE, 2011: 358-361.

[47] ALSHAMMARI R, ZINCIR-HEYWOOD A N. Can encrypted traffic be identified without port numbers, IP addresses and payload inspection?[J]. Computer networks, 2011, 55(6): 1326-1350.

[48] KORCZYŃSKI M, DUDA A. Classifying service flows in the encrypted Skype traffic[C]//Communications (ICC), 2012 IEEE International. IEEE, 2012: 1064-1068.

[49] ERMAN J, MAHANTI A, ARLITT M, et al. Semi-supervised network traffic classification[J]. ACM SIGMETRICS Performance Evaluation Review, 2007, 35(1): 369-370.

[50] XIE G, ILIOFOTOU M, KERALAPURA R, et al. SubFlow: towards practical flow-level traffic classification[C]//INFOCOM, 2012 Proceedings IEEE. IEEE, 2012: 2541-2545.

[51] HE G, YANG M, LUO J, et al. A novel application classification attack against Tor[J]. Concurrency and Computation: Practice and Experience, 2015: 27.

[52] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark[J]. ACM SIGCOMM Computer Communication Review, 2005, 35(4): 229-240.

[53] LI B, MA M, JIN Z. A VoIP traffic identification scheme based on host and flow behavior analysis[J]. Journal of Network and Systems Management, 2011, 19(1): 111-129.

[54] HURLEY J, GARCIA-PALACIOS E, SEZER S. Host-based P2P flow identification and use in real-time[J]. ACM Transactions on the Web (TWEB), 2011, 5(2): 7.

[55] SCHATZMANN D, MÜHLBAUER W, SPYROPOULOS T, et al. Digging into HTTPS: flow-based classification of webmail traffic[C]//The 10th ACM SIGCOMM Conference on Internet Measurement. ACM, 2010: 322-327.

[56] BERMOLEN P, MELLIA M, MEO M, et al. Abacus: accurate behavioral classification of P2P-TV traffic[J]. Computer Networks, 2011, 55(6): 1394-1411.

[57] XIONG G, HUANG W, ZHAO Y, et al. Real-time detection of encrypted thunder traffic based on trustworthy behavior association[M]//Trustworthy Computing and Services. Springer Berlin Heidelberg, 2013: 132-139.

[58] QIN T, WANG L, LIU Z, et al. Robust application identification methods for P2P and VoIP traffic classification in backbone networks[J]. Knowledge-Based Systems, 2015, 82: 152-162.

[59] SUN G L, XUE Y, DONG Y, et al. An novel hybrid method for effectively classifying encrypted traffic[C]//Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE. IEEE, 2010: 1-5.

[60] HE J, YANG Y, QIAO Y, et al. Fine-grained P2P traffic classification by simply counting flows[J]. Frontiers of Information Technology & Electronic Engineering, 2015, 16: 391-403.

[61] CALLADO A, KELNER J, SADOK D, et al. Better network traffic identification through the independent combination of techniques[J]. Journal of Network and Computer Applications, 2010, 33(4): 433-446.

[62] ALSHAMMARI R, ZINCIR-HEYWOOD A N. A preliminary performance comparison of two feature sets for encrypted traffic classification[C]//The International Workshop on Computational Intelligence in Security for Information Systems CISIS'08. Springer Berlin Heidelberg, 2009: 203-210.

[63] 潘吴斌, 程光, 郭晓军, 等. 基于选择性集成策略的嵌入式网络流特征选择[J]. 计算机学报, 2014, 37(10): 2128-2138.
PAN W B, CHENG G, GUO X J, et al. An embedded feature selection wsing selatine ensemble for network traffic[J]. Chinese Journal of Computers, 2014, 37(10): 2128-2138.

[64] ZHANG M, ZHANG H, ZHANG B, et al. Encrypted traffic classification based on an improved clustering algorithm[M]//Trustworthy Computing and Services. Springer Berlin Heidelberg, 2013: 124-131.

[65] DUSI M, ESTE A, GRINGOLI F, et al. Using GMM and SVM-based techniques for the classification of SSH-encrypted traffic[C]// Communications, 2009. ICC'09, IEEE International Conference. IEEE, 2009: 1-6.

[66] BAR-YANAI R, LANGBERG M, PELEG D, et al. Realtime classification for encrypted traffic[M]//Experimental Algorithms. Springer Berlin Heidelberg, 2010: 373-385.

[67] WRIGHT C V, MONROSE F, Masson G M. On inferring application protocol behaviors in encrypted network traffic[J]. The Journal of Machine Learning Research, 2006, 7: 2745-2769.

[68] WRIGHT C V, MONROSE F, MASSON G M. Using visual motifs to classify encrypted traffic[C]//The 3rd International Workshop on Visualization for Computer Security. ACM, 2006: 41-50.

[69] BONFIGLIO D, MELLIA M, MEO M, et al. Revealing skype traffic: when randomness plays with you[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(4): 37-48.

[70] WRIGHT C V, COULL S E, MONROSE F. Traffic morphing: an efficient defense against statistical traffic analysis[C]//NDSS. 2009.

[71] 何高峰, 杨明, 罗军舟, 等. Tor 匿名通信流量在线识别方法[J]. 软件学报, 2013, 24(3): 540-556.
HE G F, YANG M, LUO J Z, et al. Ouline identifrcation of Tor anongmous communication traffic[J]. Journal of Software, 2013, 24(3): 540-556.

[72] SHEN Y, LIU Y, QIAO N, et al. QoE-based evaluation model on video streaming service quality[C]//Globecom Workshops, 2012 IEEE. IEEE, 2012: 1314-1318.

[73] DERI L, MARTINELLI M, BUJLOW T, et al. nDPI: open-source high-speed deep packet inspection[C]//Wireless Communications and Mobile Computing Conference (IWCMC). IEEE, 2014: 617-622.

[74] ALCOCK S, NELSON R. Libprotoident: traffic classification using lightweight packet inspection[R]. WAND Network Research Group, Tech Rep, 2012.

[75] CARELA-ESPAÑOL V, BUJLOW T, BARLET-ROS P. Is our ground-truth for traffic classification reliable[C]//Passive and Active Measurement. Springer International Publishing. 2014: 98-108.

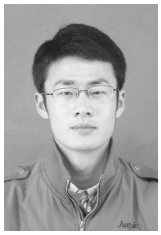
[76] GRINGOLI F, SALGARELLI L, DUSI M, et al. Gt: picking up the truth from the ground for internet traffic[J]. ACM SIGCOMM Computer Communication Review, 2009, 39(5): 12-18.

[77] QU B, ZHANG Z, ZHU X, et al. An empirical study of morphing on behavior‐based network traffic classification[J]. Security and Communication Networks, 2015, 8(1): 68-79.

[78] RAAHEMI B, ZHONG W, LIU J. Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree[C]//Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International. IEEE, 2008, 1: 525-532.

[79] ZHANG H, LU G, QASSRAWI M T, et al. Feature selection for optimizing traffic classification[J]. Computer Communications, 2012, 35(12): 1457-1471.

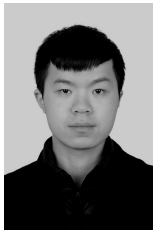
作者简介：



潘吴斌（1987-），男，江苏苏州人，东南大学博士生，主要研究方向为网络安全、网络测量及流量分类。

程光（1973-），男，安徽黄山人，东南大学教授、博士生导师，主要研究方向为网络安全、网络测量与行为学及未来网络安全。

郭晓军（1983-），男，山西长治人，东南大学博士生，主要研究方向为网络安全、网络测量及网络管理。



黄顺翔（1991-），男，湖南长沙人，东南大学硕士生，主要研究方向为网络安全、网络测量及流量分类。