文章编号:1671-4598(2021)05-0220-05 DOI:10.16526/j.cnki.11—4762/tp.2021.05.044 中图分类号:TN98,TP393.4 文献标识码:A

基于数据链路层特征的无线网络加密流量业务分类研究

尹浩东，张君毅，尚庆华
（中国电子科技集团公司第五十四研究所 河北省电磁频谱认知与管控重点实验室，石家庄 050081）

摘要：无线通信网络已经普及到人们的生活中，并且随着大众的安全意识提高，无线网络中的加密流量所占比重越来越大，网络流量加密化已成为必然趋势，其在给用户和企业带来隐私和安全的同时，也给网络安全监管和网络流量管理带来了挑战；文章研究了有线网络和无线网络的差异性，构建无线网络加密传输环境，采集无线加密网络数据，提取出数据链路层各种业务的特征并进行分类；结果表明，对不同业务的识别率在85%以上。

关键词：加密流量；无线网络；业务分类；链路层特征

Classification and Recognition of Encrypted Traffic in Wireless Networks Based on Data Link Layer Features

Yin Haodong, Zhang Junyi, Shang Qinghua
(The 54th Research Institute of CETC, Hebei Key Laboratory of Electromagnetic Spectrum Cognition and Control, Shijiazhuang 050081, China)

Abstract: The wireless communication network has spread to the life of people, and with improving the safety of the public awareness, encryption in wireless network traffic is more and more big, the network traffic encryption has become an inevitable trend, and its to the user and enterprise to bring the privacy and security at the same time, also bring to network safety and network traffic management challenges. This paper studies the difference between wired network and wireless network, constructs the wireless network encryption transmissionenvironment, collects the wireless encryption network data, extracts the data link layer each kind of service characteristic and carries on the classification. The results show that the recognition rate of different services is more than 85%.

Keywords: encrypted traffic; wireless network; service classification; link layer

0 引言

在信息化极速发展的当今社会，无线通信技术已经普及到各群众的生活中。无线技术已全面应用于商业、生活、金融及工作中。移动电话、语音通话、数字电视、网络通信、数据交换等，都随着无线通信技术的平台繁衍而生。

随着大众网络安全意识的稳步提升，对于数据保护的意识也愈加强烈。对于特定类型的流量，加密甚至已成为法律的强制性要求，数据加密俨然已经成为保护隐私的重要手段之一。根据最新统计报告截止到2019年，超过80%的企业网络流量被加密，75%的网络流量被加密。Barac预测到2020年，83%的流量将被加密。

虽然加密技术的推行旨在保护网络通信的安全和隐蔽性，但这种隐蔽性同样让它成为了攻击者隐藏部署恶意代码、渗透、命令和控制等恶意行为的强大工具。Radware公司在2016年公开的年度全球应用与网络安全研究报告中显示已有35%的恶意攻击正在借助SSL/TLS协议进行C&C命令传输、恶意代码传输等攻击活动。在2017年5月，勒索软件“想哭”(WannaCry)通过加密技术来逃避入侵检测系统的检测，致使该攻击在网络空间中如野火燎原之势传播。实现加密流量有效监管是互联网流量识别和监管的重要组成部分。加密流量识别和管理可以有效防范恶意流量，保障计算机和终端设备安全运行，维护健康绿色的网络环境。

目前国内外学者专家对于有线互联网网络加密流量分类识别的研究比较成熟，由于无线传输的介质是电磁波，相比于有线传输更易受到其他因素的干扰，而且易受相同或相近频段的无线电波影响，降低信息传输速度。因此无线通信可能引发信号损耗，降低信号传递质量，出现数据包误码和丢包问题。而且通常情况下仅能收到单向流量。

收稿日期:2021-03-01； **修回日期**:2020-03-24。
基金项目:国家自然科学基金项目(U19B2028)。
作者简介:尹浩东(1994-),男,河北石家庄人,硕士研究生,主要从事网络流量分析方向的研究。
引用格式:尹浩东,张君毅,尚庆华.基于数据链路层特征的无线网络加密流量业务分类研究[J].计算机测量与控制,2021,29(5):220-224.

因此，对于无线传输条件下网络加密流量分类识别的研究相对困难，研究成果较少。

1　无线网络与有线网络加密流传输差异性机理研究

由于有线网络与无线网络的工作频段、传输媒介等条件不同，因此在物理层和数据链路层具有差异性。但是有线网和无线网在网络层以上都遵循 TCP/IP 协议，因此在网络层以上不具有差异性。

1.1　物理层相似性与差异性

无线网络和有线网络都需要通过前导码通知设备数据链路层帧的到达。有线网络和无线网络前导码结构不同；有线网络和无线网络根据依据的标准不同，在物理层使用的技术也不同。无线网与有线网在物理层均通过前导码（preamble）通知设备数据链路层帧的到达，不同的是无线网物理层前导码包含两个部分：sync 和 SFD。其中 sync 用于发现信道中是否存在数据帧，分为长和短两个部分，一般直接称为长前导码和短前导码。其中长前导码是用于大范围低速模式，短前导码用于小范围高速模式；SFD 固定为 0000 0101 1100 1111，用作帧起始标志。对于有线以太网而言，其前导码为固定 8 个字节。根据依据的标准不同，无线网络与有线网络在物理层使用的技术也不同。

1.2　数据链路层相似性与差异性

无线网络与有线网络在数据链路层层面的差异性主要在于实现的技术与流量的传输格式。在数据链路层实现技术中，有线网络的集线器和中继器设计中采用了 CSMA/CD（Carrier Sense Multiple Access with Collision Detection，载波侦听多路访问/冲突检测）技术。该技术早期是用来解决有线网络中，共享介质下的多路网络接入问题，仍然在当今的 10M/100M 半双工网络中使用。在更高的带宽情况下，比如 1 000 M 网络，则采用全双工技术以取代 CSMA/CD。无线网络采用 CSMA/CA（Carrier Sense Multiple Access with Collision Avoidance，载波侦听多路访问/冲突避免）协议搭配停止等待协议。无线信道的通信质量远不如有线信道，因此无线站点每通过无线局域网发送完一帧后，要等到收到对方的确认帧后才能继续发送下一帧。

1.3　其他协议层次的共同性

无线网络与有线网络差异主要集中在物理层与数据链路层，在网络层以上的层次中，二者并无区别，即无线网络与有线网络均使用 TCP/IP 架构中的网络层、传输层、应用层。因此，对加密流量检测、加密流量协议分类和加密流量业务识别，在网络层以上的处理均与有线网络无异。检测要素则主要需要考虑物理层和数据链路层中的信息。

2　无线通信网络加密传输环境构建

2.1　业务环境构建

本文要对不同业务进行分类识别，因此需要构建出业务环境。目前本文选定了几种业务，分别是文件下载业务、网页浏览业务、邮件业务、即时通信业务、流媒体业务。

本文用两台计算机，一台作为服务器，一台作为客户端，建立了 FTP 服务器、邮件服务器、流媒体服务器、即时通信服务器。本文利用两台计算机进行 FTP 文件下载业务，网页浏览业务，邮件业务，流媒体业务，即时通信业务。

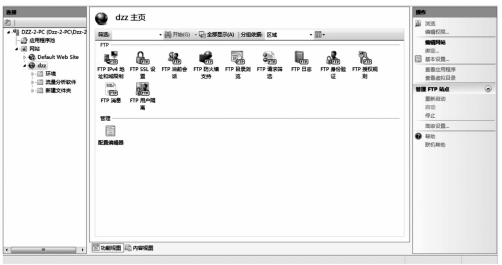

图 1　FTP 文件下载业务

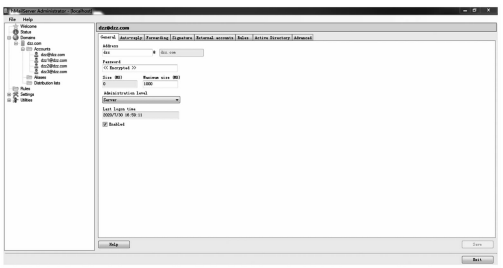

图 2　邮件业务

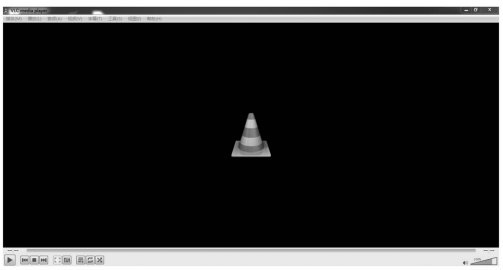

图 3　流媒体业务

图 4　即时通信业务

2.2　加密环境构建

本文利用实验室环境中的加密卫星通信系统，该系统是由一个主站和若干小站构成的，并且在数据链路层加密。作者利用其主站和两个小站，两个小站通过主站进行通信，两个小站分别连接到计算机上即可采集加密数据。结合上节构建的业务环境，就完成了加密传输的环境，为后续加密数据采集作铺垫。

3　无线网络加密数据采集

由于加密数据的隐私性及卫星通信系统的特殊性，目前没有公开的数据集，所以本文利用前文构建的加密环境

和业务环境进行加密数据的采集。

1）文件下载业务：

文件传输业务主要基于文件传输协议 FTP（file transfer protocol），它是由 TCP/IP 提供的用于从一个主机往另一个主机复制文件的标准机制。FTP 是在两个主机之间穿件了两条连接，一条用于文件传输（通常端口 20），另一条用于控制信息（通常端口 21）。在整个 FTP 回话期间，控制连接端口都是开放的，用于在客户端和服务器之间发送控制信息和客户端命令。数据连接使用的是临时端口来创建的。每当有文件要在客户端和服务器之间传输时，就创建一个数据连接。FTP 要求客户端在请求文件传输之前，发送登录名和密码给服务器，来验证自己。本文利用前文构建的业务模型中的 FTP 服务，配置好 FTP 服务器并设置好目录及文件，用另一台客户端访问 FTP 服务器地址，然后登陆 FTP 服务，访问目录并进行文件下载，并在登录服务的同时采集数据。采集几组相同文件下载的数据和几组不同文件下载的数据来做对比。

2）电子邮件业务：

邮件业务主要基于简单邮件传输协议 SMTP（simple mail transfer protocol），它是一种用于从一个服务器往另一个服务器传输的 E－mail 协议。SMTP 的特征包括邮件列表、回复接收和转发。SMTP 可以接收输入的消息，并利用 TCP 把它发送给另一个服务器上的 SMTP。SMTP 的作用是利用本地电子邮件数据包把输入消息存储在用户的收件箱中。一旦 SMTP 服务器标志出了接收者的 E－mail 服务器的 IP 地址，就将通过标准的 TCP/IP 路由过程发送消息。本文利用构建好的电子邮件服务，分别在连接到两个小站的计算机上登录配置好的邮件账号，开启数据采集设备，两个账户之间互相发文字信息，互相传送附件并下载。

3）流媒体业务：

流媒体业务主要基于实时传输协议 RTP（Real－time transport protocol），它用来为网络上的语音、图像、传真等多种需要实时传输的多媒体数据提供端到端的实时传输服务。RTP 既不需要实现建立连接，也不需要中间节点的参与。在网络带宽充足的情况下，RTP 具有一定的带宽调控能力，保证端到端的多媒体流同步。在网络带宽不足时，RTP 的带宽调控能力将受到一定的限制。本文利用构建好的流媒体服务器 VLC，在服务器端配置好串流视频属性，在客户端配置好串流地址，开始进行数据采集。采集几组同样视频的数据和不同视频的数据来作对比。

4）即时通信业务：

即时通信业务主要是为用户提供即时消息，语音，视频，文件传输等多样化服务。即时通信业务是一种基于 Internet 的通信技术，涉及到 IP/TCP /UDP 等多种技术手段。无论即时通信系统的功能多么复杂，它们大都基于相同的技术原理，主要包括客户/服务器（C/S）和对等通信（P2P）模式。本文利用 FeiQ 在两台计算机之间进行通信，采集文字、图片等消息的发送接收和文件的传送等数据。

4 无线通信网络分层加密数据的特征选择及提取

4.1 无线网络分层加密数据特征差异性研究

1）无线加密流量物理层与数据链路层特征：

无线网络与有线网络的主要差异在物理层和数据链路层中，因此本文对于无线加密流量进行特征提取的研究点也主要集中在物理层和数据链路层的协议特征提取中。

相较于有线信道，无线信道为了保证数据传输的安全性，有些情况在数据链路层就进行了加密处理。但是，与 TLS 等安全传输层加密协议类似，无线网络两个通信节点建立连接时，经历了 802.11 相互发现过程、802.1X 认证过程和 4 次握手过程，这些过程中会包含大量的伴生明文信息。而在正常通信过程中数据链路层中还存在一些未被加密的数据帧字段，这均可以作为无线加密流量的特征进行识别。由于这种加密通信大部分情况下需要手动进行配置（例如在路由器设置中手动开启使用 WPA2），这也有可能导致无线信道中可能存在未经加密的报文，灵活运用这些未被加密的报文，可以较为方便的对网络层及以上的特征进行提取。

2）无线加密流量的网络层与传输层特征：

由于无线网络与有线网络在网络层及以上并无明显差异，因此，在可以完整提取到网络层及以上报文的前提条件下，无线网络与有线网络的网络层与传输层特征并无太大差异。但是，如果使用 TCP 协议作为传输层协议以实现数据的严格按序传输，相较于有线网络，无线网络环境将面临三点主要的问题：1）由于信号衰减等多种问题，无线信道的丢包率明显较高；2）无线信道是不对称的，主要体现在带宽不对称、丢包率不对称与路由不对称 3 个方面，这将导致测量结果产生偏差，进而无法正确设置 TCP 重传定时器的超时时间；3）由于通信范围的有限，无线网络存在隐患终端和暴露终端问题，这将导致时隙资源的无序争用，增加了报文碰撞的概率，进而增大了数据传输时延，严重影响网络的吞吐量。因此无线网络中存在比有线网络更多的重传报文，在对会话进行特征提取时需要对这些情况进行特别的识别与处理。

4.2 特征选择与提取方法研究

为了进行识别与分析，需要对无线加密流量的特征进行提取，本论文拟通过对无线网络流量进行分析，归纳出无线加密流量的特征池，为后续的研究奠定基础。

从无线信道中抓取到的数据帧有可能从数据链路层开始就已经得到加密处理，这样将无法正常对网络层及以上的特征进行获取，这对加密流量识别产生了很大困扰。因此相较于有线网络加密流量识别特征提取主要提取网络层及以上特征，针对无线加密流量的特征提取来源更加广泛。

由于有些无线通信系统物理层就是加密的，所以无法提取到上层的特征，必须对物理层的特征进行分析。通过对采集的大量数据的十六进制数据流进行分析，可以通过不加密的帧头分离出业务帧与控制帧。

图 5 是对文件下载业务的帧长统计，文件下载业务主

要分为两部分完成，包括 FTP 的登录和文件的传输，FTP 登录时客户端请求访问需要将自己的登录名和密码发送给服务器来验证。从图中可以看出，0～50 帧左右帧长度在 100～600 字节小幅波动，此时为 FTP 的登录过程；在后续 50～1 700 帧为文件的传输过程，可以看出此时帧长度基本可以保持在最大值 1 500 字节，并且比较稳定。

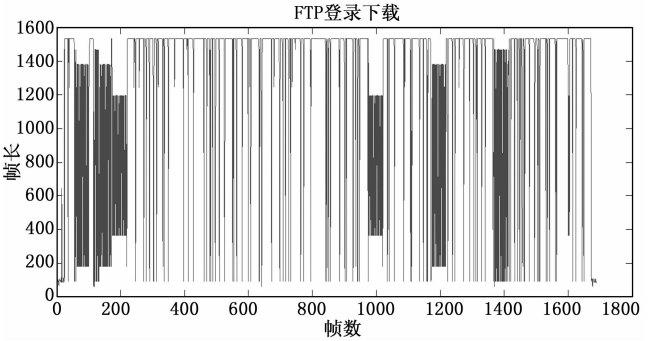

图 5　文件下载业务帧长统计

图 6 是对流媒体业务的帧长统计，从图中可以看出，流媒体业务的帧长波动范围很大，从 100～1 500 字节均有分布，但是基本都在 200 字节以上由图可知，流媒体业务的帧长度波动幅度很大，最大帧长度可达到 1 500 字节。可以看出流媒体业务帧长波动幅度较大，最大帧长可达到 1 525 字节。

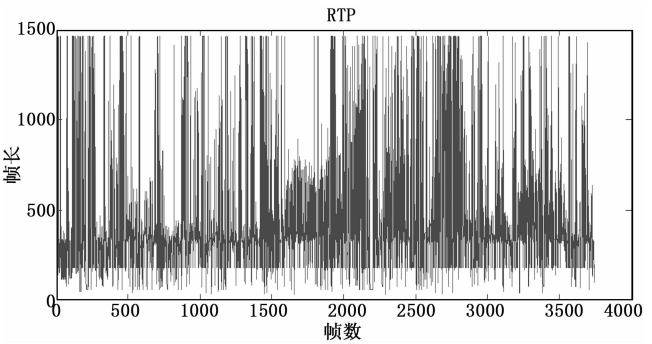

图 6　流媒体业务帧长统计

图 7 是对邮件业务的帧长统计，从图中可以看出，在 0～30 帧左右帧长在 100～600 字节波动，在 30～120 帧左右在 100 字节左右波动较小，在 120～1 800 帧左右在 100～1 500 字节波动但 1 500 字节占很大比例。通过分析发现，在 0～30 帧左右为邮件的登录过程，在 30～120 帧左右为邮件发送文字业务，在 120～1 800 帧左右为邮件发送附件的业务，此时和文件下载业务类似。

图 8 是对即时通信业务的帧长统计，从图中可以看出，在 0～20 帧左右帧长为 100 字节左右，可能为通信双方交互过程；在 20～30 帧左右有一个较大值，可达到 1500 字节；在 30～100 帧基本维持在 100 字节左右，可能是保持通信的数据帧；在 100～150 帧有一段 1 500 字节的峰值，可能是消息通信；在 150～350 帧也有一段 1 500 字节的数据帧，可能是消息通信或者文件传输；在 560～860 帧有比较多的 1 500 字节的数据帧，可能是文件传输过程。

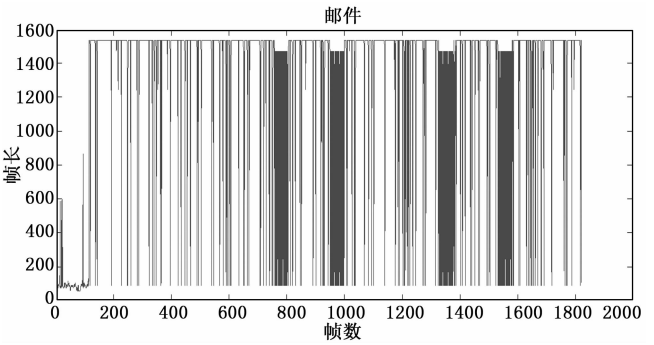

图 7　邮件业务帧长统计

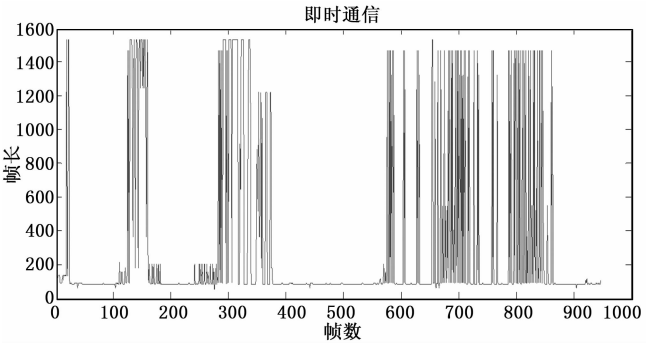

图 8　即时通信业务帧长统计

5　无线网络加密流业务分类

5.1　算法介绍

本论文利用 K－means 算法进行特征分类，该算法核心是通过设定参数作为个子集的中心点，将计算数据集中的点与中心点的相似性，将点归入相似性最高的子集中，然后在每个子集中计算均值选择中心点，重复以上步骤直至中心点不再变化。其中计算相似性使用最小化平方差来计算：

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} | dis(k_j, c_j) |^2 \qquad (1)$$

其中：E 为数据集中所有点之间的均方差之和，x_j 为随机选择的数据集中非本轮中心点的某一点，c_j 为本轮选择的中心点，K－means 算法是基于参数 K 预先设定，并且受包含与正常值差异较大的噪声数据影响较大，算法的具体步骤如下：

输入：聚类个数 K 和具有 n 个对象的数据集。

输出：K 个聚类中心点及其对象。

1）在包含 n 个对象的数据集中随机选取 K 个对象作为中心点；

2）计算与中心点的距离，将数据集中剩余数据对象聚到与之距离最小的中心点的类簇中；

3）在每个类簇中重新计算得到 n 个中心点；

4）重复步骤 2）和 3）直至中心点不再发生变化；

5）输出结果。

5.2　业务分类结果及分析

本文利用前文构建的加密传输环境得到的加密数据，结合上文提取的特征和方法，对文件下载业务、即时通信业务、流媒体业务和邮件业务进行分类识别，得到的结果如图 9 所示。可以看出，每种业务的分类识别率都在 85％以上。

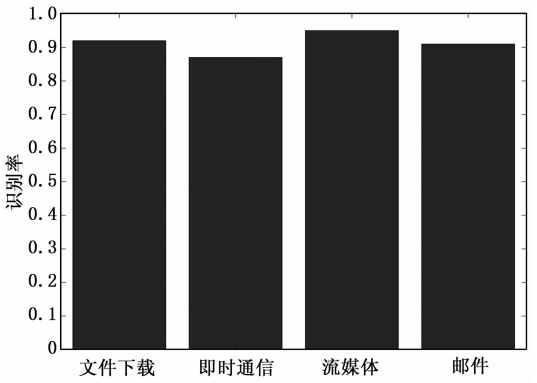

图 9 业务分类识别率

现有针对网络加密流量分类识别的研究主要是研究网络层及以上加密数据，本文研究的是数据链路层加密数据。通过对 4 种业务的加密流量分类结果分析得出，本文提出的无线网络数据链路层加密流量特征也可以对业务进行分类，并且由于加密层次在数据链路层，对于数据的要求更加广泛，适用范围更广。但是加密层次低带来的一个问题是数据中包含的信息相对于网络层及以上包含的信息较少，由此带来对于某些业务例如即时通信业务的某些特征会与其他业务的相似性较大造成识别率相对较低。本文进一步的研究方向是对流量特征进行更深层次的挖掘以寻找更多可用于分类的特征。

6 结束语

目前国内外学术界对专门无线通信网络加密流的测量与识别还是一片空白，其主要研究着眼于无线通信网络加密技术、无线网络测量技术和非网络环境相关的加密流量识别技术。国内除近几年兴起的加密流量识别与分析领域的研究外，其他领域的研究相较于国外而言相对落后。因此，国内亟需对无线通信加密网络的安全通信进行研究，而无线通信网络加密流的测量与识别则可以为未来无线通信加密网络的安全通信奠定基础。

本文创新性地面向无线通信数据准确识别的需求，考虑到实际无线通信网络环境中存在的问题，研究针对无线通信网络加密流的测量与识别技术，突破目前国内在无线通信网络加密流测量与分析领域的空白，打破国际在该领域的技术垄断，实现对无线网络的有效监管，并反哺推动无线通信加密技术的发展，保障我国未来无线网络通信的安全，为国家网络安全保驾护航。

参考文献：

[1] 赵 彩，丁 凰. 网络信息安全中 DES 数据加密技术研究 [J]. 计算机测量与控制，2017，25 (8)：241 - 247.

[2] 王 影. 无线网络流量异常数据信息检测仿真 [J]. 计算机仿真，2017，34 (9)：408 - 411.

[3] 潘吴斌，程 光. 网络加密流量识别研究综述及展望 [J]. 通信学报，2016，37 (9)：154 - 167.

[4] 于 强，霍红卫. 一组提高存储效率的深度包检测算法 [J]. 软件学报，2011，22 (1)：149 - 163.

[5] 徐 鹏，林 森. 基于 C4. 5 决策树的流量分类方法 [J]. 软件学报，2009，20 (10)：149 - 163.

[6] 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.

[7] 马若龙. 基于卷积神经网络的未知和加密流量识别的研究与实现 [D]. 北京：北京邮电大学，2018.

[8] 余金澳，吴彦鸿. 一种面向方位敏感性的 PCA - SVM 分类识别方法 [J]. 无线电工程，2017，47 (5)：87 - 90.

[9] 高长喜，吴亚飚，王 枞. 基于抽样分组长度分布的加密流量应用识别 [J]. 通信学报，2015，36 (9)：65 - 75.

[10] 鲁信金，施育鑫，雷 菁. 基于遗传算法的全索引调制的物理层加密算法 [J]. 无线电通信技术，2020，46 (2)：173 - 179.

[11] 谢希仁. 计算机网络第五版 [M]. 北京：电子工业出版社，2008.

[12] Balachandran A，Voelker G M，Bahl P，et al. Characterizing user behavior and network performance in a public wireless LAN [A]. Proceedings of the 2002 ACM SIGMETRICS international conference on Measurement and modeling of computer systems [C]. 2002，195 - 205.

[13] Chinchilla F，Lindsey M，Papadopouli M. Analysis of wireless information locality and association patterns in a campus [A]. IEEE INFOCOM 2004 [C]. IEEE，2004，2：906 - 917.

[14] Balachandran A，Voelker G M，Bahl P，et al. Characterizing user behavior and network performance in a public wireless LAN [A]. Proceedings of the 2002 ACM SIGMETRICS international conference on Measurement and modeling of computer systems [C]. 2002，195 - 205.

[15] Yeo J，Youssef M，Agrawala A. A framework for wireless LAN monitoring and its applications [A]. Proceedings of the 3rd ACM workshop on Wireless security [C]. 2004，70 - 79.

[16] Banerjee S，Agrawala A K. Estimating available capacity of a network connection [A]. Proceedings IEEE International Conference on Networks 2000 (ICON 2000). Networking Trends and Challenges in the New Millennium [C]. IEEE，2000，131 - 138.

[17] Chinchilla F，Lindsey M，Papadopouli M. Analysis of wireless information locality and association patterns in a campus [A]. IEEE INFOCOM 2004 [C]. IEEE，2004，2：906 - 917.

[18] Velan P. A survey of methods for encrypted traffic classification and analysis [J]. International Journal of Network Management，2015，25 (5)：355 - 374.

[19] Dorfinger P，Panholzer G，John W. Entropy estimation for real－time encrypted traffic identification (short paper) [A]. International Workshop on Traffic Monitoring and Analysis [C]. Springer，Berlin，Heidelberg，2011：164 - 171.

[20] 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.