综述

移动设备网络流量分析技术综述

徐明[1]，杨雪[1,2]，章坚武[1]

（1. 杭州电子科技大学网络空间安全学院，浙江 杭州 310018;
2. 浙江警察学院计算机与信息技术系，浙江 杭州 310053）

摘　要：移动设备在人们的日常生活中不可或缺，分析移动设备产生的网络流量能够为网络管理、隐私保护等活动提供有价值的信息。为深入了解流量分析在移动设备领域的发展现状及趋势，介绍了网络流量分析的基本框架和移动设备网络流量的收集手段，并分类总结了移动设备网络流量分析的目的。最后根据目前相关研究中仍存在的问题，对移动设备网络流量分析领域的研究方向进行了展望。

关键词：移动设备；网络管理；隐私保护；网络流量分析

中图分类号：TP393

文献标识码：A

doi: 10.11959/j.issn.1000-0801.2018151

A review of network traffic analysis targeting mobile devices

XU Ming[1], YANG Xue[1,2], ZHANG Jianwu[1]

1. School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China
2. Department of Computer and Information Technology, Zhejiang Police College, Hangzhou 310053, China

Abstract: Mobile devices are now ubiquitous in people's life, thus analyzing network traffic generated by these devices can provide valuable information for network management and privacy preservation. In order to review the works that contribute to the state of the art of network analysis targeting mobile devices, the framework of network analysis, traffic collection approaches were introduced and a classification of the works according to the goal of the analysis was presented. Finally, the future research of network analysis targeting mobile devices was proposed based on the current problems in this literature.

Key words: mobile device, network management, privacy preservation, network traffic analysis

1　引言

随着社会信息化、网络化大潮的推进，移动设备（如智能手机和平板电脑等）越来越多地渗透到人们的日常工作与生活中，成为全球数十亿人不可获取的工具。根据数据互联网统计公司

收稿日期：2018−01−06；**修回日期**：2018−04−11

基金项目：网络空间安全重点专项基金资助项目（No.2016YFB0800201）；国家自然科学基金资助项目（No.61572165，No.61702150）；浙江省自然科学基金重点资助项目（No.LZ15F020003）；浙江省重点研发计划基金资助项目（No. 2017C01065，No. 2017C01062）

Foundation Items: Cyberspace Security Major Program in National Key Research and Development Plan of China (No.2016YFB0800201), The National Natural Science Foundation of China (No. 61572165, No. 61702150), The State Key Program of Zhejiang Province Natural Science Foundation of China (No.LZ15F020003), Key Research and Development Plan Project of Zhejiang Province (No.2017C01065, No. 2017C01062)

Statista 的统计，2016 年，全球智能手机用户总人数为 21 亿，预计这一数值将于 2020 年增长至 28.7 亿[1]。与传统网络流量相比，Wi-Fi 网络的广泛部署以及应用市场中大量可用的应用程序使得移动设备不仅能够保障传统的通信活动（如拨打语音电话、发送短消息），还更多地应用于金融、在线游戏和网络购物等高级场景。移动设备中往往存储其持有者的隐私数据（如联系人、照片、视频以及 GPS 位置等），因而越来越多的攻击者及流量分析人员瞄准其产生的网络流量，试图从中挖掘有用的信息。

网络流量分析是计算机信息安全领域的一个分支，它将一组设备产生的网络流量作为输入，以与这些设备、用户、应用程序或流量本身有关的信息作为输出。如图 1 所示，网络流量分析通常包括 4 个阶段：流量收集、预处理、数据分析、结果评估。流量收集是构建数据集的过程；预处理则通过去除数据集中无效的数据或提取流量的关键特征等将收集到的数据转换为可理解的格式以便后续分析。数据分析是网络流量分析流程中最重要的环节，可根据所采用的技术分为以下 4 类。

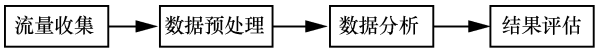

图 1　网络流量分析流程

（1）基于端口的分析方法

基于端口的分析方法是最古老的网络流量分析技术，具体做法为从 TCP/UDP 报文头部提取端口号，并与互联网数字分配机构（Internet Assigned Numbers Authority，IANA）为各类应用分配的 TCP/UDP 端口列表进行对比，从而推断网络流量的来源。基于端口的方法容易实现，且不受流量加密的影响，常应用于防火墙或访问控制列表。然而，基于端口的网络流量分析技术容易被端口混淆、NAT（network address translation）、端口转发及随机端口分配等技术影响，采用基于端口的网络流量分析技术在分类流量时正确率不足 70%[2]。

（2）深度报文检测方法

深度报文检测（deep packet inspection，DPI）技术分析应用层的网络流量，通过事先提取各应用程序的模式，利用提取到的模式在未知网络流量中区分不同流量的来源。由于应用程序的底层逻辑可能会随时间推移发生变化，因而 DPI 技术需要定期更新提取到的模板。此外，DPI 技术往往受到网络流量加密措施的影响。

（3）基于图的分析方法

从网络流量中构建并分析不同的应用程序或主机的交互图，并应用于应用程序分类。Karaginanis 等人[3]最早利用图来表示主机在应用层的交互模式并构建图样本库，然后从未知流量中构建图并去样本库中匹配，从而实现对应用程序的识别。

（4）基于统计和机器学习的分析方法

基于统计和机器学习的流量分析方法是近年来应用最多的方法，该方法假定不同应用程序产生的网络流量具有独特的统计分布特性，在已知网络流量中训练获得分布特性并应用于分析未知网络流量。在应用此类方法时，如何妥善选取报文特征及统计工具（或机器学习算法）是非常重要的因素，直接影响分析结果的正确率。Ferreira 等人[4]采用元分析的手段调研了 2005—2017 年网络流量分析领域的主要文献，为从事网络流量分析研究的同行在网络报文特征选择方面提供了有效的建议。在机器学习算法的选择方面，常用的分类器有随机森林、决策树、高斯朴素贝叶斯和支持向量机等。

网络流量分析流程的最后一个阶段是结果评估，依照标准评估此前所采用的分析方法是否理想，常用的评价标准包括 TP（true positive）、FN（false negative）、TN（true negative）、FP（false positive）、精确率、召回率及 F1 值等。

移动设备网络流量分析由传统的网络流量分析领域发展而来，与传统网络流量相比，移动设备产生的下载流量比上载流量大[5-7]，流量持续时间较短、报文数目较多且单个报文长度较短[8]。大多数应用层流量通过 HTTP 或 HTTPS 协议传递[6,8-13]，但使用 HTTPS 传输应用层流量是未来发展的趋势。视频流贡献了移动设备网络流量中很重要的一部分[11,14]。Android 与 iOS 操作系统上免费应用程序中嵌入的广告及定位服务触发了大量的网络流量[15]。本文介绍了常用的移动设备网络流量的收集方法，并根据移动设备网络流量分析的目标对现有研究进行分类综述，本文中的网络流量指由移动设备产生的互联网流量，不包含蜂窝数据流量、蓝牙数据等其他类型的数据。

2 网络流量收集

移动设备产生的网络流量可从网络各个层次（如数据链路层、传输层、应用层）或节点（Wi-Fi 网络接入点或设备内）收集，数据来源主要有：移动设备、网络访问点（access point，AP）、Wi-Fi 监控器和运行移动设备仿真器的计算机。

最直接的流量收集方法是在移动设备上安装轻量级的应用记录程序。除此之外，在可控制的小规模网络中，也可利用小型网关、VPN 服务器和台式计算机等作为 AP 来记录用户的网络流量，图 2 展示了这种流量收集方式。

随着移动用户对 Wi-Fi 网络需求的增长，Wi-Fi 访问点也被用于流量收集。Wi-Fi 网络通常包括两种类型的硬件设备：采用 IEEE 802.11 标准为移动设备提供网络连接的 AP 与将来自 AP 的网络流量转发至互联网的网关。Wi-Fi 调制解调器等硬件既可用作 AP 又可用作网关。在 Wi-Fi 访问点进行流量收集又分为网络中只存在单访问点和网络中存在多访问点的情况。

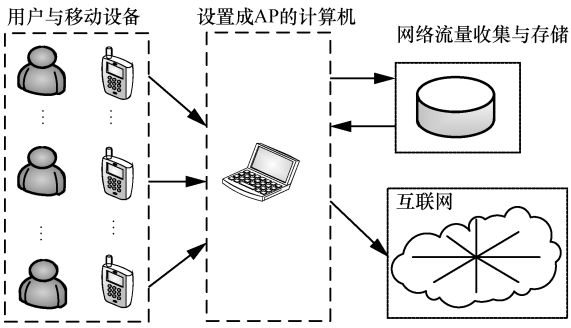

图 2 将计算机设置为 AP 以收集流量

Wi-Fi 监听器是一种能够扫描 Wi-Fi 频段以获取网络流量的硬件设备。常见于将传统的 Wi-Fi 设备（如 PCI 卡或台式计算机）设置为监听模式，使其被动监听附近的 Wi-Fi 信号。为了有效地监听 Wi-Fi 设备的网络流量，该监听设备必须处于目标网络的覆盖范围内，其有效性受到选取的监听频段、Wi-Fi 调制解调器的频率以及周围建筑物等因素的影响。

移动设备仿真器是能够虚拟设备组件及操作系统的虚拟机，是一种应用程序测试方案。由于运行该仿真器的计算机负责在仿真器和互联网间转发网络流量，因此该计算机可作为网络流量的理想收集点。

3 移动设备网络流量分析的研究目的

移动设备网络流量分析的目的是从网络流量中推断设备、用户及设备上安装的应用程序的相关信息。图 3 展示了对这一领域研究目的的分类。

3.1 设备信息推断

移动设备上运行的操作系统类型及版本是重要的设备相关信息。操作系统识别指通过分析网络流量判断移动设备运行的操作系统。攻击者在识别目标设备的操作系统后可进一步定制后续的攻击，而在人群聚集的情况下，操作系统识别技术可服务于市场或社会调研。

Coull 与 Dyer 等人[16]针对苹果公司的即时通信工具 iMessage，利用用户与苹果服务器间交换的加密报文长度来确定用户使用的是 iOS 或 OSX

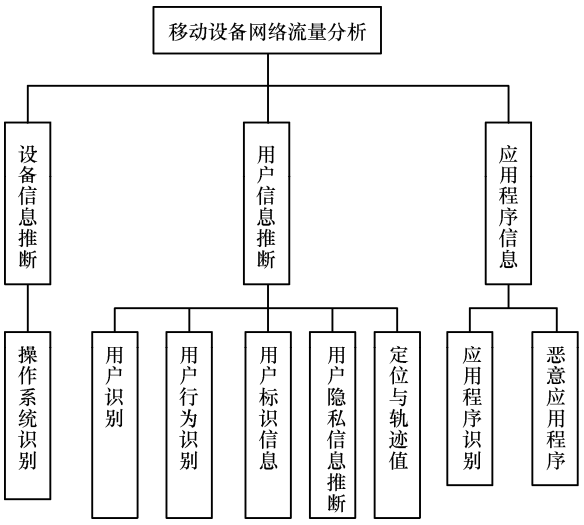

图 3　移动设备网络流量分析研究目的分类

操作系统，方法不受流量加密的影响，且仅需观察流量中的前 5 个报文就能够达到 100%的识别率。

Ruffing 等人[17]认为移动设备产生的网络流量的时序特征由该移动设备运行的操作系统决定，提出通过分析网络报文时序的频谱以识别与操作系统相关的频率特性。由于无需审查报文内容，该方法可适用于流量加密的场景。收集运行 Android、iOS、Windows Phone 和 Symbian 等操作系统的智能手机产生的流量，在观察时长为 5 min 的报文序列后检测率能够达到 90%。然而该方法在判断同一操作系统的不同版本时效果不理想。

Malik 等人[18]利用移动设备产生的报文时间间隔来推断该设备运行的操作系统并成功区分 Android、iOS 与 Windows Phone。由于只使用报文时间间隔信息，该方法不受流量加密的影响。然而，在实验阶段仅使用了 3 台设备，即每台设备运行一种操作系统平台。因此，无法判断同一操作系统的不同版本对检测率的影响。

3.2　用户信息推断

（1）用户识别

用户通过移动设备接入互联网，使用应用程序。同一用户在不同移动设备、不同网络中产生的网络流量模式基本是稳定的。通过学习用户的网络模式可以推断某些网络流量是否来自特定用户。

Vanrykel 等[19]开发了一款自动执行应用程序并收集网络流量的工具，审查 HTTP 数据并识别明文传送的敏感标识符，利用这些敏感标识符提取来自某特定用户的网络流量。他们收集了来自 42 个类别的 1 260 个 Android 应用程序产生的流量，并成功将同一用户的 57%的流量关联起来。然而，由于该方法需要使用 HTTP 报文内容，因此无法处理加密后的流量。

Verde 等[20]使用互联网提供商从用户网络流量中提取出的统计数据“NetFlow 记录”训练分类器，能够精确识别某一用户是否连接到指定网络并获取其 IP 地址，且不受 NAT 技术影响。他们收集了通过同一 Wi-Fi 接入点连接互联网的 26 个用户的网络流量，选取随机森林作为分类器获得 95%的正确率和 7%的误判率。值得说明的是，由于“NetFlow 记录”不包含网络报文数据，因此该方法不受流量加密的干扰。

（2）用户行为识别

用户在使用应用程序时会执行若干操作并触发网络数据传输。这些操作涉及“用户—应用程序”间的交互，因而某一特定的操作会呈现固定的模式（如浏览 Facebook 个人主页所产生的流量与在 Twitter 上发送推文的流量模式不同），这些模式可被用于在网络流量中识别用户特定的行为。用户行为识别可用于市场或调查分析，还可用来实现去匿名化。目前，用户行为识别针对的应用程序大多属于即时通信类（iMessage、WhatsApp）、社交（Facebook、Twitter）、邮件客户端（Gmail、Yahoo Mail）等。

Coull 和 Dyer 等人[16]检测用户在 Apple 即时通信工具 iMessage 上执行的“开始输入”“停止输入”“发送信息”“发送附件”“阅读”5 个动作，并在 iOS 系统上获得 99%的正确率。除此之外，

他们还对用户使用的自然语言（中文、英文、法文、德文、俄文和西班牙文）所交互的信息长度进行推测。由于该方法选取用户与 Apple 服务器间交换报文的长度为特征，因此他们的方法不受信息加密的影响。

Park 和 Kim[21]研究韩国即时聊天应用程序 KakaoTalk 支持的 11 种用户行为（加入聊天室、发送信息、添加朋友等），得到每种行为模式对应的报文特征序列，该序列被用于在未知网络流量中识别对应的用户行为。由于选取报文长度为特征，因此该方法在 KakaoTalk 流量加密的情况下仍能得到 99.7%的准确率。

Shafiq 等人[22]首次提出识别微信用户发送文本及图片产生的流量，在其所处校园及宿舍收集了两个网络流量数据集，从中提取 44 个统计特征，采用 C4.5、贝叶斯网络、支持向量机和朴素贝叶斯 4 种分类器对用户发送文本和图片所产生的流量进行分类。实验结果显示，C4.5 和支持向量机对这两种用户行为所产生的流量的分类效果最好，分别能够达到 99.91%和 99.57%的精确率。

Conti 等人[23]使用 IP 地址、TCP 报文头部信息识别用户在同一应用中执行的不同操作。其核心思想为当用户在同一应用中执行不同操作时，其出站和入站网络流的特征不同。由于以 TCP 报文流中出站及入站字节数序列作为特征统计量，因而适用于在传输层加密的流量。与参考文献[23]类似，Fu 等人[24]提出识别用户在同一应用内不同操作的系统 CUMMA，他们先将收集到的流量划分为 Session 和 Dialog 两个层次，以 Dialog 为单位选择报文长度序列和时间间隔序列作为特征，Wechat 和 WhatsApp 两款应用（包含 8 种不同用户操作）上的实验表明，使用随机森林作为分类器时 CUMMA 系统的整体正确率能够达到 96%以上。

（3）用户标识信息检测

移动设备上安装的应用程序大多请求访问用户的敏感信息，如 GPS 定位、照片、联系人等，并需要接入互联网。个人标识信息（personal identifiable information，PII）可用于识别、定位用户。移动设备中通常包含以下 4 种 PII。

- 移动设备相关信息，如国际移动设备识别号（international mobile equipment identity，IMEI）、Android 设备 ID（Android 设备第一次启动时随机生成的标识符）与 MAC 地址（网卡的唯一标识符）等。
- SIM 卡相关信息，如 IMSI（international mobile subscriber identity）与 SIM 序列号。
- 用户信息，如姓名、性别、出生日期、居住地址、电话号码和电子邮件地址等。
- 用户地理位置信息，如 GPS 定位和邮政编码等。

由于移动设备上安装的应用程序更容易产生易被识别的流量特征，因此，在移动设备上使用应用程序比使用浏览器访问相同的服务更容易泄露隐私数据[25]。个人标识信息检测技术通过分析网络流量检测是否存在敏感信息泄露。

Stevens 等人[7]研究 Android 系统 13 个广告库，通过分析广告网络流量以检测是否有用户标识信息泄露。他们发现在 2012 年仅有 1 个广告库提供商对其网络流量实现了加密。对广告库触发的网络流量执行深度报文检测，显示多种用户标识信息（如年龄、性别、GPS 位置）被以明文的方式泄露。参考文献[7]指出，即使广告库提供商并未刻画用户画像，外部攻击者仍可利用其网络流量中泄露的唯一设备标识符（unique device identifier，UDID）从不同的广告库提供商关联用户的敏感信息并构建完整的用户画像。

Kuzuno 与 Tonami [26]调查免费 Android 应用程序中加载的广告库对用户敏感信息的泄露情况，关注移动设备标识符、IMSI 与 SIM 序列号以及运营商等信息。以多款泄露用户敏感信息的应用程序产生的流量为输入，对这些数据进行聚类以生成流量签名，然后使用这些签名检测移动设

备上其他应用程序是否会泄露用户敏感信息。由于该方法需要审查 HTTP 报文以获得签名，因此无法应用于加密后的网络流量。

跨平台的系统 Recon[27]能够使用户控制移动设备产生的网络流量对其敏感信息的泄露。Recon 基于跨平台的网络流量收集与分析系统 Meddle[28]，由于 Meddle 利用 VPN 隧道将目标设备的流量重定向到自己的代理服务器，因此能够处理在传输层加密的流量。此外，Recon[27]还提供网页接口使用户实时查看被泄露的敏感信息，并选择修改这些信息或阻止泄露敏感信息的网络连接。

AntMonitor[29]是一个收集并分析 Android 设备流量的系统，它在收集到的流量数据中查找 IMEI、Android ID、手机号码、电子邮件地址和设备定位等用户标识信息，检测用户标识信息是否遭到泄露。由于 AntMonitor 在执行用户标识信息检测时查看应用层数据，因此无法应用于加密流量。

开源 Android 应用程序 PrivacyGuard[30]利用 Android API 的 VPNService 类窃听安装在 Android 设备上的应用程序产生的网络流量。PrivacyGuard 被用来监测应用程序是否泄露与用户（如电话号码）或设备（如 IMEI）相关的敏感信息。与前人的研究成果 TaintDroid[31]相比，PrivacyGuard 能够检测出更多被泄露的信息且能够伪造数据替换它们。PrivacyGuard 能够利用中间人技术处理在传输层加密的网络流量。

Continella 等人[32]开发开源工具 Agrigento 分析 Android 应用程序以检测用户标识信息是否遭到泄露。首先在设备上多次运行 Agrigento 收集设备产生的网络流量及系统和应用级信息（如随机生成的标识符、时间戳等）；然后对移动设备操作系统中的敏感信息赋一些特殊值，再次运行该工具收集网络流量和执行时信息。若前后收集到的数据模型不一致，则提示敏感信息泄露。与 Recon[27]相比，Agrigento 在控制误报率的同时，能够检测到更多被泄露的敏感信息。与 PrivacyGuard 类似该工具也能够使用中间人技术查看 HTTPS 报文信息，处理在传输层加密的流量。

（4）用户隐私信息推断

目前，Wi-Fi 网络数据泄露用户隐私的问题已获得广泛关注，但从 Wi-Fi 流量元信息及移动设备属性（如安装的应用程序、连接的 Wi-Fi 网络）中推断用户的社会属性也是一个需要重视的威胁。

Barbera 等人[33]研究能否从大量移动设备发送的 Wi-Fi 探测请求中推断用户的社会属性。他们在商场、火车站和大学校园等区域收集超过 16 万设备发送的 11 MB 探测数据，生成移动设备用户社交图，该社交图中用户及服务集标识（service set identifier，SSID）等属性呈现幂次现象。参考文献[33]得出用户之间关系越紧密越倾向于选择同一厂商的移动设备的结论，并利用设备提供商识别号推断用户年龄及社会地位。

Li 等人[34]提出一个无需检查 Wi-Fi 流量内容推断用户隐私信息的系统 DIP。该系统使用流量中的 MAC 地址、IP 地址作为位置特征，HTTP 报文头部的 host 及 user-agent 字段作为应用特征，采用随机森林模型作为分类器，在大学校园网络内推断用户的性别及教育水平（博士研究生、硕士研究生、本科生），并分别达到 78%和 74%的精确率。在流量加密的情况下，由于无法获取 host 和 user-agent 特征，其精确率受到影响降为 67%和 72%。此外，还讨论了使用 Tor 网络、MAC 地址随机化、随机发送无效报文等技术防御隐私推断攻击。

（5）定位与轨迹估计

用户频繁访问的地点通常揭示其社会地位、兴趣及爱好等隐私，这些信息可用于商业用途（如定点投放广告）或警方的侦查活动。定位目标用户的移动设备是一种简单有效地获取此类信息的方式。设备定位指通过移动设备产生的网络流量

来推断其地理位置，而轨迹估计指通过分析移动设备产生的网络流量，推断其在某地理区域内的移动轨迹。通过轨迹估计能够获得单个用户的兴趣、社会习性等，也可以聚合多个用户的轨迹来预测路网的交通状况。

恶意网络（即由一组恶意 Wi-Fi 设备构成的网络）能够定位移动设备的地理位置[35]，网络中的每个恶意 Wi-Fi 节点都会查找包含目标设备 MAC 地址的 Wi-Fi 查询请求报文，并将其所得信息上传至中央服务器。中央服务器利用其从多个节点获得的信息进行分析，从而定位目标设备。参考文献[35]中利用软件仿真城市中携带启用了 IEEE 802.11 协议的移动设备用户，研究结果表明采用最新的 IEEE 802.11 标准更容易构建定位移动设备地理信息的网络。

移动设备周期性发送的 Wi-Fi 查询请求可被用于设备跟踪[36]。攻击者可通过部署多个 Wi-Fi 监听器将监测到的 Wi-Fi 查询请求发送给中央服务器，中央服务器将多个监听器对同一移动设备的监测结果聚合为一条时空轨迹。部署了 3 个监听器并通过 GPS 信息作为位置监测对照数据来评估系统。实验结果表明，当所部署的 Wi-Fi 监听器相互距离为 400 m 时，该系统的平均地理位置定位误差在 70 m 以内。

3.3 应用程序信息推断

（1）应用程序识别

应用程序的网络流量指纹指由应用程序产生的网络流量表现出的模式，可用来在未知网络流量中识别应用。应用程序识别对提高网络服务质量（quality of service，QoS）或网络安全防护等有积极意义。

Lee[37]等人选取 Android 和 iOS 应用市场排名前 50 的应用生成网络流量指纹，利用这些指纹在未知网络流量中检测是否存在相关应用。实验结果显示，与已收集到的指纹匹配的网络流量仅占测试流量的 15.37%，这一结果表明智能手机用户在应用程序的使用方面存在很大差异。由于使用最长公共子序列（LCS）从 HTTP 报文中提取指纹，因此该方法无法处理加密后的网络流量。

SAMPLES 是一个自适应的应用程序识别框架[38]，把 HTTP 流量中应用程序标识符的出现模式抽象为包含应用标识符的 HTTP 字段、标识符的前缀及后缀字符串的一组文本规则。利用“聚合—验证”的方式提高规则的精确度并确定优先级，并将调优后的规则加载到应用程序识别引擎中，通过“提取—查找”的模式识别网络流量对应的应用程序。为了验证 SAMPLES 的有效性，选取了 70 万个 Android 应用程序，并收集了 1 500 万网络流，实验结果表明 SAMPLES 能够识别出 90%的应用程序，并能够达到 99%的正确率。然而，由于文本规则来源于 HTTP 报文，因此 SAMPLES 对加密后的流量无效。

Wang 等人[25]指出即使应用程序网络流量加密，攻击者也可以利用边通道泄露的信息识别用户的行为。他们将应用程序接收及发送的流量划分为多个子序列，并提取序列中各个报文长度、到达时间、传输方向以及报文长度平均值、标准差等共 20 项统计值作为特征，使用随机森林算法识别应用并达到很好的效果。此外，还指出由于移动设备上的应用程序产生的流量模式更易被识别，因而用户使用移动设备访问服务比使用浏览器访问同一服务更易泄露信息。

AppScanner[39]从应用程序接收、发送及双向网络流中提取 54 种统计特征，训练随机森林分类器识别未知网络中的应用程序，并评估流量收集时间、移动设备、操作系统版本及应用程序版本等因素对检测率的影响。AppScanner 从 TCP 和 IP 报文头部提取特征，因而能够处理被 SSL/TLS 加密的流量。Alan 等[40]利用 Android 应用程序启动阶段产生的网络流量的 TCP/IP 报文头部信息识别应用程序，与参考文献[39]类似，参考文献[40]同样评估了流量收集时间、移动设备、操作系统及

运营商对应用识别正确率的影响，并得出操作系统、运营商信息发生变化时对结果影响最大，而测试集与训练集间隔数天对检测结果几乎无影响。

（2）恶意应用程序检测

移动设备中往往包含大量用户敏感信息，很容易成为恶意软件开发者攻击的对象，因而应用程序市场、安全公司以及移动用户对恶意软件的检测技术非常关注。

Su[41]等人提出一个由部署在云上的验证服务器和 Android 设备构成的恶意应用程序检测框架供 Android 应用市场使用。开发者在应用软件发布前需要将其提交至验证服务器检测，Android 设备则用于执行开发者提交的程序并监控系统调用和网络流量。服务器基于系统调用统计和网络流量特征判断应用程序恶意与否，采用随机森林作为网络流量分类器达到 96.7%的正确率。

Wei[42]等人提出一种 Android 恶意软件检测框架，利用恶意 Android 应用程序产生的流量学习它们的行为。除此之外，该框架能够自动分析某指定应用程序的 DNS 流量并判断是否恶意。使用 Android 恶意软件公开数据集和从 Android 应用市场收集的正常应用程序对框架进行评估，正确率、召回率和精确率接近 100%。由于该框架需要访问应用程序生成的 DNS 流量，因而无法应用于加密网络流量。

参考文献[43]利用恶意应用程序通常会将用户敏感信息发送到恶意远程主机这一特点，记录移动设备上安装的应用程序与远程主机的全部通信，并利用已知的恶意域名，将与恶意域名主机交互的应用程序标记为恶意。同样，该方法需要访问应用程序产生的 HTTP 报文中的 URL，因此无法适用于加密流量。

Narudin 等人[44]研究基于恶意的入侵检测系统能否依靠流量分析检测恶意 Android 应用程序。他们通过在移动设备上运行正常应用程序，在动态分析平台上运行恶意应用程序来收集网络流量数据集，并将这些数据发送到训练多款分类器的中央服务器。该方法使用了 HTTP 报文中的信息，因而无法处理加密后的网络流量。

TrafficAV 是一款基于机器学习的 Android 恶意软件检测系统，能够分别提供基于 TCP 和 HTTP 报文特征的检测模型[45]。由于 HTTP 流量能够提供更加丰富的信息，因而基于 HTTP 报文特征的检测模型检测率高于基于 TCP 报文特征的模型检测率，但基于 HTTP 报文特征的模型无法处理加密后的网络流量。

Arora 等人[46]收集恶意软件流量的样本，从网络层提取特征（如接收报文的平均时间间隔、网络流发送的字节数等）训练朴素贝叶斯分类器，并使用与训练集不同的、来自 6 个家族的恶意应用程序评估其检测方法。此外，还提出一种特征选择算法，在保证检测率不会骤降的情况下，减少所使用的特征数量。该方法的优势在于不受网络流量加密的影响。

Shabtai 等人[47]设计了一款基于主机的 Android 恶意应用程序检测系统，通过监控设备的内存、网络、电量等提取特征，训练多种分类器（决策树、贝叶斯网络等）检查设备是否安装了恶意应用程序，并评估当测试应用程序未被用于训练及训练和测试阶段在不同移动设备上执行等情形对检测结果的影响。他们还设计了一款基于恶意行为的 Android 恶意应用程序检测系统[48]，识别恶意的攻击及设备上安装的看似“正常”，实则是注入了恶意代码后重新打包的应用程序。此外，针对当年 Google 官方应用市场出现的具有自动更新能力的恶意应用程序，提出基于网络流量模式的检测办法，学习正常应用程序的网络流量模式与待测应用程序表现出的流量模式比较，判断待测应用程序是否恶意。实验结果表明，应用程序在感染恶意代码前后表现出明显不同的流量模式，因而恶意应

用在运行数分钟后就能被检测到。

4 结束语

移动设备网络流量分析是现阶段的研究热点。本文结合近年来的相关研究成果，归纳了目前常用的流量收集方式，并从设备、用户及应用程序 3 个方面对流量分析的研究目的进行分类。从研究目的来看，目前用户识别、应用识别、隐私信息泄露检测等方向研究成果较多，而用户定位、轨迹估计和隐私信息推断等方向研究相对较少。从流量分析的网络层次来看，与低层（网络层及传输层）相比，高层（应用层）网络流量提供更加丰富的信息，往往能够得到更好的结果，但低层网络流量分析更适用于流量加密的情况，因此研究者应关注如何在加密流量中挖掘更多有用信息。此外，海量的网络数据是现有网络安全分析机制面临的一大挑战[49-52]，除目前常用的机器学习方法以外，如何在保证流量分析效果的同时，压缩流量特征、减少待处理的数据[53]也将是未来研究的方向。

参考文献：

[1] STATIST A. Number of smartphone user worldwide from 2014 to 2020 (in billions) [EB]. 2016.

[2] MOORE A, PAPAGIANNAKI K. Toward the accurate identification of network applications [C]//International Conference on passive and active network measurement, March 31-April 1, 2005, Boston, MA, USA. Heidelberg: Springer-Verlag, 2005: 41-54.

[3] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark [C]//ACM Special Interest Group on Data Communication, August 22-26, 2005, Philadelphia, PA, USA. New York: ACM Press, 2005: 229-240.

[4] FERREIRA D, VAZQUEZ F, VORMAYR G. A meta-analysis approach for feature selection in network traffic research [C]// The Reproducibility Workshop, August 21-25, 2017, Los Angeles, NV, USA. [S.l.:s.n.], 2017.

[5] LINDORFER M, NEUGSCHWANDTNER M, WEICHSELBAUM L, et al. ANDRUBIS – 1,000,000 apps later: a view on current Android malware behaviors[C]// The 3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, September 11, 2014, Wroclaw, Poland. Washington: IEEE Computer Society, 2014: 3-17.

[6] SHEPARD C, RAHMATI A, TOSSELL C, et al. LiveLab: Measuring wireless networks and smartphone users in the field [J]. ACM SIGMETRICS Performance Evaluation Review, 2010, 38(3): 15-20.

[7] STEVENS R, GIBLER C, CRUSSELL J, et al. Investigating user privacy in Android Ad libraries[C]//The 2012 Workshop on Mobile Security Technologies, May 24, 2012, San Francisco, CA, USA. [S.l.:s.n.], 2012.

[8] CHEN X, JIN R, SUH K, et al. Network performance of smart mobile handhelds in a university campus WI-FI network[C]// The 2012 ACM SIGCOMM Internet Measurement Conference, August 13-16, 2012, Helsinki, Finland. New York: ACM Press, 2012: 315-328.

[9] CHEN Z, HAN H, YAN Q, et al. A first look at Android malware traffic in first few minutes[C]//2015 IEEE Trustcom/BigDataSE/ISPA, August 20-22, 2015, Helsinki, Finland. New York: IEEE Computer Society, 2015: 206-213.

[10] NAYAM W, LAOLEE A, CHAROENWATANA L, et al. An analysis of mobile application network behavior[C]//The 12th Asian Internet Engineering Conference, November 30-December 2, 2016, Bangkok, Thailand. New York: ACM Press, 2016: 9-16.

[11] GEMBER A, ANAND A, AKELLA A. A comparative study of handheld and non-handheld traffic in campus Wi-Fi networks[C]//The 12th International Conference on Passive and Active Measurement, March 20-22, 2011, Atlanta. Heidelberg: Springer-Verlag, 2011: 173-183.

[12] WEI X, VALLER N, MADHYASTHA H, et al. Characterizing the behavior of handheld devices and its implications [J]. Computer Networks, 2017(114): 1-12.

[13] FALAKI H, LYMBEROPOULOS D, MAHAJAN R, et al. A first look at traffic on smartphones[C]//The 2010 ACM SIGCOMM Internet Measurement Conference, November 1-3, 2010, Melbourne, Australia. New York: ACM Press, 2010: 281-287.

[14] AFANASYEV M, CHEN T, VOELKER G, et al. Usage patterns in an urban Wi-Fi network[J]. IEEE/ACM Transactions on Networking, 2010, 18(5): 1359-1372.

[15] ESPADA A, GALLARDO M, SALMERON A, et al. Performance analysis of Spotifyc for Android with model-based testing [J]. Mobile Information Systems, 2017.

[16] COULL S, DYER K. Traffic analysis of encrypted messaging services: Apple iMessage and beyond [J]. ACM SIGCOMM Computer Communication Review, 2014, 44(5): 5-11.

[17] RUFFING N, ZHU Y, LIBERTINI R, et al. Smartphone reconnaissance: Operating system identification[C]//13th IEEE Annual Consumer Communications and Networking Conference, January 9-12, 2016, Las Vegas, NV, USA. New York: IEEE

Communications Society, 2016: 1086-1091.

[18] MALIK N, CHANDRAMOULI J, SURESH P, et al. Using network traffic to verify mobile device forensic artifacts[C]// The 14th IEEE Annual Consumer Communications and Networking Conference, January 8-11, 2017, Las Vegas, NV, USA. Piscataway: IEEE Press, 2017: 114-119.

[19] VANRYKEL E, ACAR G, HERRMANN M, et al. Leaky birds: exploiting mobile application traffic for surveillance [C]// the 20th International Conference on Financial Cryptography and Data Security, February 22-26, 2016, Barbados. Heidelberg: Springer-Verlag, 2017: 367-384.

[20] VERDE N, ATENIESE G, GABRIELLI E, et al. No NAT'd user left behind: fingerprinting users behind NAT from NetFlow records alone[C]//The 34th IEEE International Conference on Distributed Computing Systems, June 30-July 3, 2014, Madrid, Spain. Washington: IEEE Computer Society, 2014: 218-227.

[21] PARK K, KIM H. Encryption is not enough: inferring user activities on KakaoTalk with traffic analysis[C]//The 16th International Workshop on Information Security Applications, August 20-22, 2015, Jeju Island, Korea. Heidelberg: Springer-Verlag, 2015: 254-265.

[22] SHAFIQ M, YU X Z, LOGHARI A, et al. WeChat text and picture messages service flow traffic classification using machine learning technique[C]//The 14th International Conference on Smart City, December 12-14, 2016, Sydney, Australia. [S.l.: s.n.], 2016.

[23] CONTI M, MANCINI L, SPOLAOR R, et al. Analyzing Android encrypted network traffic to identify user actions [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(1): 114-25.

[24] FU Y J, XIONG H, LU X J, et al. Service usage classification with encrypted Internet traffic in mobile messaging apps [J]. IEEE Transactions on Mobile Computing, 2016, 15(11): 2851-2864.

[25] WANG Q L, YAHYAVI A, KEMME B, et al. I know what you did on your smartphone: inferring app usage over encrypted data traffic[C]//The 2015 IEEE Conference on Communications and Network Security, September 28-30, 2015, Florence, Italy. New York: IEEE Communications Society, 2015: 433-441.

[26] H. KUZUNO AND S. Tonami. Signature generation for sensitive information leakage in Android applications[C]//The 29th IEEE International Conference on Data Engineering, April 8-12, 2013, Brisbane, Australia. Washington: IEEE Computer Society, 2013: 112-119.

[27] REN J, RAO A, LINDORFER M, et al. ReCon: revealing and controlling PII leaks in mobile network traffic[C]//The 14th Annual International Conference on Mobile Systems, Applications, and Services, June 26-30, 2016, Singapore. New York: ACM Press, 2016: 361-374.

[28] RAO A, MOLAVI KAKHKI A, RAZAGHPANAHS A, et al. Using the middle to meddle with mobile [R]. 2012.

[29] LE A, VARMARKEN J, LANGHOFF S, et al. AntMonitor: a system for monitoring from mobile devices[C]//2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data, August 17-21, 2015, London, UK. New York: ACM Press, 2015: 15-20.

[30] SONG Y, HENGARTNER U. PrivacyGuard: a VPN-based platform to detect information leakage on Android devices[C]// The 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices, October 12, 2015, Denver, USA. New York: ACM Press, 2015: 15-26.

[31] ENCK W, GILBERT P, CHUN B G, et al. TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones[C]//The 9th USENIX Symposium on Operating Systems Design and Implementation, October 4-6, 2010, Vancouver, Canada. Berkeley: USENIX Association, 2010: 393-407.

[32] CONTINELLA A, FRATANTONIO Y, LINDORFER M, et al. Obfuscation-resilient privacy leak detection for mobile apps through differential analysis[C]//The 2017 Network and Distributed System Security Symposium, February 26-March 1, 2017, San Diego, USA. Reston: Internet Society, 2017.

[33] BARBERA M, EPASTO A, MEI A, et al. Signals from the crowd: Uncovering social relationships through smartphone probes[C]//The 2013 ACM SIGCOMM Internet Measurement Conference, October 23-25, 2013, Barcelona, Spain. New York: ACM Press, 2013: 265-276.

[34] LI H , XU Z , ZHU H , et al. Demographics Inference through WI-FI network traffic analysis [C]//The 35th IEEE International Conference on Computer Communications, April 10-15, 2016, San Francisco, USA. Piscataway: IEEE Press, 2016: 1-9.

[35] HUSTED N, MYERS S. Mobile location tracking in metro areas: Malnets and others[C]//The 17th ACM Conference on Computer and Communications Security, October 4-8, 2010, Chicago, USA. New York: ACM Press, 2010: 85-96.

[36] MUSA A, ERIKSSON J. Tracking unmodified smartphones using Wi-Fi monitors[C]//The 10th ACM Conference on Embedded Networked Sensor Systems, November 6-9, 2012, Toronto, Canada. New York: ACM Press, 2012: 281-294.

[37] LEE S, PARK J, LEE H, et al. A study on smart-phone traffic analysis[C]//The 13th Asia-Pacific Network Operations and Management Symposium, September 21-23, 2011, Taipei, China. New York: IEEE Communications Society, 2011: 177-183.

[38] YAO H, RANJAN G, TONGAONKAR A, et al. SAMPLES: self adaptive mining of persistent LExical Snippets for classifying mobile application traffic[C]//The 21th Annual International Conference on Mobile Computing and Networking, September 7-11, 2015, Paris, France. New York: ACM Press, 2015: 439-451.

[39] TAYOR V, SPOLAOR R, CONTI M, et al. AppScanner: automatic fingerprinting of smartphone Apps from encrypted net-

work traffic [C]//The 1st IEEE European Symposium on Security and Privacy, March 21-24, 2016, Saarbrucken, Germany. Piscataway: IEEE Press, 2016: 439-454.

[40] ALAN H, KAUR J. Can Android applications be identified using only TCP/IP headers of their launch time traffic[C]//The 9th ACM Conference on Security and Privacy in Wireless and Mobile Networks, July 18-20, 2016, Darmstadt, Germany. New York: ACM Press, 2016: 61-66.

[41] SU X, CHUAN M, TAN G. Smartphone dual defense protection framework: detecting malicious applications in Android markets[C]//The 8th International Conference on Mobile Ad-hoc and Sensor Networks, December 14-16, 2012, Chengdu, China. Washington: IEEE Computer Society, 2012: 153-160.

[42] WEI T E, MAO C H, JENG A B, et al. Android malware detection via a latent network behavior analysis[C]//The 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, June 25-27, 2012, Liverpool, UK. Washington: IEEE Computer Society, 2012: 1251-1258.

[43] ZAMAN M, SIDDIQUI T, AMIN M, et al. Malware detection in Android by network traffic analysis[C]//The 1st International Conference on Networking Systems and Security, January 5-7, 2015, Dhaka, Bangladesh. Piscataway: IEEE Press, 2015: 1-5.

[44] NARUDIN F, FEIZOLLAH A, ANUAR N, et al. Evaluation of machine learning classifiers for mobile malware detection [J]. Soft Computing, 2016, 20(1): 1-5.

[45] WANG S, CHEN Z, ZHANG L, et al. TrafficAV: an effective and explainable detection of mobile malware behavior using network traffic[C]//The 24th IEEE/ACM International Symposium on Quality of Service, June 20-21, 2016, Beijing, China. Piscataway: IEEE Press, 2016: 384-389.

[46] ARORA A, PEDDOJU S. Minimizing network traffic features for Android mobile malware detection[C]//The 18th International Conference on Distributed Computing and Networking, January 4-7, 2017, Hyderabad, India. New York: ACM Press, 2017: 32.

[47] SHABTAI A, KANONOV U, ELOVICI Y, et al. “Andromaly”: a behavioral malware detection framework for Android devices [J]. Journal of Intelligent Information Systems, 2012, 38(1): 161-190.

[48] SHABTAI A, TENENBOIM-CHEKINA L, MIMRAN D, et al. Mobile malware detection through analysis of deviations in application network behavior [J]. Computers & Security, 2014, 43(6): 1-18.

[49] 汪来富, 金华敏, 刘东鑫, 等. 面向网络大数据的安全分析技术应用[J]. 电信科学, 2017, 33(3): 112-118.
WANG L F, JIN H M, LIU D X, et al. Application of security analysis technology for network big data[J]. Telecommunications Science, 2017, 33(3): 112-118.

[50] 姜红红, 张涛, 赵新建, 等. 基于大数据的电力信息网络流量异常检测机制[J]. 电信科学, 2017, 33(3): 134-141.
JIANG H H, ZHANG T, ZHAO X J, et al. A big data based flow anomaly detection mechanism of electric power information network[J]. Telecommunications Science, 2017, 33(3): 134-141.

[51] 王帅, 汪来富, 金华敏, 等. 网络安全分析中的大数据技术应用[J]. 电信科学, 2015, 31(7): 145-150.
WANG S, WANG L F, JIN H M, SHEN J, et al. Big data application in network security analysis [J]. Telecommunications Science, 2015, 31(7): 145-150.

[52] 曹旭, 曹瑞彤. 基于大数据分析的网络异常检测方法[J]. 电信科学, 2014, 30(6): 152-156.
CAO X, CAO R T. Network anomaly prediction method based on big data [J]. Telecommunications Science, 2014, 30(6): 152-156.

[53] NASR M, HOUMANSADR A, MAZUMDAR A. Compressive traffic analysis: a new paradigm for scalable traffic analysis [C]// The 2017 ACM Conference on Computer and Communications Security, October 30-November 3, 2017, Dallas, USA. New York: ACM Press, 2017: 2053-2069.

[作者简介]

徐明（1970−），男，博士，杭州电子科技大学网络空间安全学院教授、博士生导师，主要研究方向为网络信息安全。

杨雪（1988−），女，浙江警察学院计算机与信息技术系讲师，杭州电子科技大学计算机学院博士生，主要研究方向为网络信息安全。

章坚武（1961−），男，博士后，杭州电子科技大学教授、博士生导师，主要研究方向为移动通信与个人通信。