软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(10):3205−3222 [doi: 10.13328/j.cnki.jos.005620]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

流量混淆技术及相应识别、追踪技术研究综述*

姚忠将[1,2], 葛敬国[1], 张潇丹[1], 郑宏波[1], 邹 壮[1,2], 孙焜焜[1,2], 许子豪[1,2]

[1](中国科学院 信息工程研究所,北京 100093)

[2](中国科学院大学 网络空间安全学院,北京 100049)

通讯作者: 张潇丹, E-mail: zhangxiaodan@iie.ac.cn

摘 要: 流量混淆技术是目前审查规避系统常用技术之一.为了提升网络流量识别精度和监管能力,针对混淆流量的识别和追踪技术也备受关注.通过深入分析随机化、拟态和隧道这 3 类主流的流量混淆技术,对比了其技术框架、隐蔽性、易用性和应用场景;分析了深度包检测、机器学习等两类识别技术,对比了其识别精度;分析对比了被动关联、主动关联两类流量追踪技术.最后给出了流量混淆、识别和追踪技术的发展趋势.

关键词: 流量混淆;VPN;Tor;流量识别;流量追踪

中图法分类号: TP309

中文引用格式: 姚忠将,葛敬国,张潇丹,郑宏波,邹壮,孙焜焜,许子豪.流量混淆技术及相应识别、追踪技术研究综述.软件学报,2018,29(10):3205−3222. http.//www.jos.org.cn/1000-9825/5620.htm

英文引用格式: Yao ZJ, Ge JG, Zhang XD, Zheng HB, Zou Z, Sun KK, Xu ZH. Research review on traffic obfuscation and its corresponding identification and tracking technologies. Ruan Jian Xue Bao/Journal of Software, 2018,29(10): 3205−3222 (in Chinese). http.//www.jos.org.cn/1000-9825/5620.htm

Research Review on Traffic Obfuscation and Its Corresponding Identification and Tracking Technologies

YAO Zhong-Jiang[1,2], GE Jing-Guo[1], ZHANG Xiao-Dan[1], ZHENG Hong-Bo[1], ZOU Zhuang[1,2], SUN Kun-Kun[1,2], XU Zi-Hao[1,2]

[1](Institute of Information Engineering, The Chinese Academy of Science, Beijing 100093, China)

[2](School of Cyber Security, University of Chinese Academy of Science, Beijing 100049, China)

Abstract: Traffic obfuscation technology is one of the most commonly used techniques in censorship-circumvention systems. In order to improve the recognition accuracy and supervisory ability of network traffic, much attention has been paid to the recognition and tracking of obfuscated traffic. Through in-depth analysis of three main traffic confusion technologies, such as randomization, mimicry and tunneling, this paper compares the technical framework, concealment, ease of use and application scenarios of the traffic confusion technologies. In addition, the paper reviews two types of recognition technology: deep packet inspection and machine learning, and compares their recognition accuracy. Furthermore, it analyzes and compares two types of traffic tracing technology: passive and proactive correlation. Finally, it discusses the identification and trace technology development trends of obfuscation traffic.

Key words: traffic confusion; VPN; Tor; traffic identification; traffic tracing

审查规避系统(censorship-circumvention system)是协助互联网用户绕过网络审查的流量伪装技术统称,包括匿名通信网和虚拟专用网(virtual private network,简称 VPN)技术等.在匿名通信网技术方面,最早的审查规避系统采用 Chaum[1]于 1981 年提出的 Mix 技术,通过中间节点扰乱审查者的视线.随后出现了 Anonymizer[2]、

*基金项目: 国家重点研发计划(2017YFB0801801); 国家科技重大专项(2017ZX03001019-003)

Foundation item: National Key R&D Plan of China (2017YFB0801801); National Science and Technology Major Project (2017ZX0300 1019-003)

收稿时间: 2018-01-23; 修改时间: 2018-04-16; 采用时间: 2018-06-04

Crowds[3]、DC-Net[4]、P5[5]、I2P[6]等匿名通信网.目前应用最广泛、关注度最高的是美国海军创建的第一个低时延匿名通信网 Tor(the onion routing).截至 2017 年 11 月 1 日,Tor 的全球 Tor relay 用户达到 3 000 000,使用传输插件(pluggable transport,简称 PT)的 Tor Bridge 用户有 43 000.在虚拟专用网方面,VPN 因其部署简单、性能较高等特点而被广泛应用.目前市场上 VPN 产品种类繁杂,国外知名的有 lantern[7]、Psiphon[8]等 30 多款,国内有 Snap VPN[9]、极速安全 VPN[10]、蝙蝠 VPN[11]等 40 多款.据著名市场研究机构 Global WebIndex 2017 年调查报告估测,亚洲 VPN 用户占上网用户的 31%[12],中国 VPN 用户多达 9 000 万[13].

审查规避系统若被用于从事商业犯罪和政治犯罪等活动会给社会造成严重的负面影响.据中国互联网违法和不良信息举报中心资料显示,利用境外服务器、VPN 等网络资源向中国境内网民实施网络犯罪已成为当前网络犯罪的突出动向[14].ISIS 恐怖分子经常使用匿名通信网络内运行的 Mail2Tor 和 SIGAINT 邮件服务工具,通常难以追踪[15].为了规范网络空间秩序,加强网络安全治理,很多国家、组织和公司纷纷出台审查制度,部署相关审查系统.2012 年,Twitter 宣布将根据不同国家要求部署用户信息审查技术[16].2017 年,我国工信部发布了《工业和信息化部关于清理规范互联网网络接入服务市场的通知》,在全国范围内对互联网非法网站和 VPN 开展清理规范工作[17].相应地,审查规避系统也得以迅速发展.

为了提高审查规避能力,审查规避系统利用流量混淆技术将非正常流量隐藏于正常流量中,难以区分.审查规避系统通常在接入匿名通信网的第 1 跳或连接 VPN 代理节点之前引入混淆技术.流量混淆技术的不断升级也增强了审查规避系统的抗审查能力.以 Tor 匿名通信网络为例,它以传输插件的形式将混淆技术集成到 Tor 浏览器,将 HTTP 报文混淆处理后发送出去.VPN 则在 VPN 客户端将报文混淆处理后发往 VPN 代理节点.早期审查规避系统简单地依赖加密报文负载隐藏信息,但是审查者依靠 IP 地址、服务端口号等特征可以轻易识别[18].为此,审查规避系统依靠加密、转换、填充等随机化方法来隐藏指纹信息、长度分布等特征,如 Dust[19]、Obfs[20−22]、Scramble Sui[23]等.考虑到随机化方法难以抵御基于熵测试和启发式检测的组合攻击,有研究者提出了协议拟态技术,通过流量整形使非正常流量具备普通流量的指纹、格式等特征,如 FTE[24]伪装成基于 HTTP 的密文格式、CensorSpoofer[25]模拟加密的 VoIP 会话、SkypeMorph[26,27]模仿 Skype 视频流量等.即便如此,审查者仍可通过统计分析报文中的 URL 熵值或长度特征识别拟态混淆流量.隧道技术是一种更强的流量混淆技术,直接将非正常数据加密封装进普通协议报文中,达到规避审查的目的,如 CloudTransport[28]、Meek[29]、Decoy-Routing[30].研究发现,隧道技术可以依靠流量分析等技术加以识别.本文深入分析了流量混淆技术,总结其混淆框架并分析其相应的隐蔽性.

对混淆流量的识别技术按照识别特征和方法分为深度包检测技术和基于机器学习的流量识别技术.深度包检测技术针对 3 种混淆技术分别总结相应识别技术:随机化混淆流量的识别方法主要依靠某字段或报文的熵作为识别依据(如一条 Obfs 流前 2048 字节的熵[31]);拟态混淆流量的识别方法可以依靠识别特征字符(如 StegoTorus 传输 PDF 文件时的 xref 关键字[31])和某字段或报文的熵(如 FTE URI 的熵[31]);隧道混淆流量的识别方法比较丰富,包括基于协议字段(如 Meek 的 TLS 加密套件[32])、基于长度(如基于 SSL 的 Tor 报文长度[18])、基于熵(如 Meek 报文间隔分布相对熵[33])、基于行为模式(如 Tor 的 Circuit 建立过程状态转换[34]).本文将基于机器学习的流量识别技术按 3 类混淆技术进行划分:随机化混淆流量和拟态混淆流量可采用 *k*NN、朴素贝叶斯和 CART 算法利用已知特征识别(如 Obfs、FTE);隧道混淆流量可采用 SVM、聚类、决策树等机器学习算法依据提取特征识别和深度神经网络直接依据流量数据识别(如 Meek、Tor).本文分析了深度包检测、机器学习两类流量识别技术根据混淆特征识别混淆流量,对比相互间识别精度并汇总了混淆流量识别技术.

流量追踪技术可以进一步地发现非正常流量的发送者和接收者.流量追踪技术分为被动关联技术和主动关联技术.被动关联技术通过对采集的流量进行分析来关联具有相似特征的流量,达到流量追踪的目的.虽然操作简单,但是数据采集量大,计算开销大.主动关联技术包括流水印技术和渗透技术.流水印技术在疑似混淆流量中嵌入标记信息并在潜在接收端检测流量,如果检测出标记信息,表明追踪成功.主要包括基于流速特征(如 DSSS[35])和时间特征(如 ICBW[36])的流水印嵌入方法.虽然简单、高效,但不同的流水印技术抵御丢包、乱序、篡改等干扰的能力参差不齐,对多流攻击(multi-flow attack,简称 MFA)[37]和均方自相关攻击(mean-square

autocorrelation attack,简称 MSAC)[38]的抵抗力较弱.渗透技术应用广泛,包括中间人攻击、节点发现和重放攻击3类.准确率较高,但是部署难度大,成本较高.

本文第1节深入分析随机化、拟态和隧道这3类流量混淆技术,对比其技术框架、隐蔽性、易用性和应用场景.第2节分析深度包检测、机器学习这两类识别技术,对比其识别精度.第3节分析对比被动关联、主动关联两类流量追踪技术.第4节给出流量混淆、识别和追踪技术的发展趋势.

1 流量混淆技术

网络监管需识别网络中的流量,依据流量类型调配网络资源、限制流量传输等.识别流量的依据就是流量指纹.本文首先定义流量指纹和流量混淆技术.

定义 1(流量指纹). 流量指纹是表征某一流量的一个特征或一系列的特征组合,包括静态指纹特征(如字段信息、报文长度等特征字段信息)和动态指纹特征(如熵值、报文长度分布等统计特征).

定义 2(流量混淆). 任何可将目标流量置于观测流量集中无法识别的状态均称为流量混淆.

混淆技术的目的就是隐藏流量指纹特征,避免基于深度包检测技术的审查.目前常见的流量混淆技术按实现原理分为3类[31,39]:(1) 随机化(randomizer);(2) 拟态(mimicry);(3) 隧道(tunneling).本节将分析3类混淆技术的实现框架,探讨每种混淆技术的典型案例并分析每类混淆技术的隐蔽性.

1.1 随机化

定义 3(随机化流量混淆). 利用加密、随机填充、随机时延调整、位运算等方法随机化目标流量特征字段、字符和部分流量统计特征等信息,使观察者难以从观测流量集中识别目标流量的状态称为随机化流量混淆.

用以随机化混淆流量的技术称为随机化流量混淆技术,旨在隐藏非正常流量的静态指纹特征和部分动态特征.通过分析常用随机化混淆工具,本文总结了随机化混淆技术的通用框架,介绍了随机化的典型实例.

1.1.1 随机化混淆框架

随机化混淆技术在发送端、接收端分别部署调制器、解调器.调制器和解调器通常作为调制解调模块集成在客户端和服务端.调制器负责随机化运算,解调器负责随机化逆运算.随机化与逆随机化运算可以形式化地描述为

$$P' = Random(P,E,F,A,S,B),\ P = Random^{-1}(P',E^{-1},F^{-1},A^{-1},S^{-1},B^{-1}),$$

其中,P'是随机化报文,$Random()$为随机化运算,$Random()^{-1}$为$Random()$的逆运算,E为加密参数,E^{-1}为解密参数,F为填充参数,F^{-1}为去填充参数,A为报文间隔调整参数,A^{-1}为去报文间隔调整参数,S为分割参数,S^{-1}为合并参数,B为位运算参数,B^{-1}为逆位运算参数.对两种运算来说,只有待处理报文和加密参数是必需的.

图1所示为随机化混淆技术通信过程:客户端发送报文P,经过调制器转换为P',P'经审查网络到达解调器,解调器将P'逆随机化还原P并转发给服务端.

Fig.1 Architecture of random obfuscation technology

图1 随机化混淆技术框架

1.1.2 典型实例

Brandon提出的Dust[19]采用带外半握手技术(out-of-band half-handshake technique)协商密钥,提出3种类型报文(Invite、Intro和Data报文)均携带MAC、IV字段.MAC由密文、IV和随报文类型变化的密钥计算得出.IV是加密密文和计算MAC的一次性随机值.除MAC、IV及随机填充字段外,其余字段均被加密.Dust可有效规避基于静态指纹和部分动态指纹特征的深度包检测,对基于报文间隔特征的检测无能为力.

Kadianakis 和 Mathewson 提出的 Obfs2[20]是基于 Brl 的 SSH 混淆协议,在密钥协商阶段,依据随机长度数 PADLEN(范围 0~MAX_PADDING)决定密钥填充以随机化密钥长度,通信阶段以加密方法随机化报文负载,但未随机化报文长度.Obfs2 采用的 Diffie-Hellman(DH)公钥很容易与同样大小的随机字符串区分,两人又提出了 Obfs3[21].Obfs3 使用的 UniformDH 公钥与统一的 1 536 比特字符串相比,可忽略字符长度的不同,同时提出另一个随机数 PADLEN2(范围 0~MAX_PADDING/2)填充报文.Philipp 等人[23]为提高 Obfs3 抗主动攻击的能力和混淆流量指纹,提出 ScrambleSuit.ScrambleSuit 从所有报文长度分布中随机选取一种报文长度填充为不小于 MTU 的长度;用共享的 PRNG 种子生成伪随机分布,决定 Bin 数量后为每个 Bin 指定[0,10]范围内的概率,从分布中抽取报文间隔调整随机值;提出 protocol polymorphism 降低协议分类精度.Angel 和 Winter 基于 ScrambleSuit,提出 Obfs4[22].Obfs4 在握手报文负载填充随机长度以混淆初始流签名,完成握手后将应用层数据拆分成“packets”加密传输.为避免识别长度字段,帧长度通过与 OFB 模式的 SipHash-2-4 做 XOR 运算.Obfs4 对报文全部加密,造成在普通报文应是明文的字段被随机化,审查者可据此识别 Obfs4 混淆流量.

1.2 拟 态

定义 4(拟态流量混淆). 利用正则表达式转换、借用连接等方法,辅以加密、填充等技术,将目标流量特征整形为样本流量特征,使目标流量难以从观测流量集识别的状态称为拟态混淆.

用以模拟样本流量特征的技术称为拟态流量混淆技术,旨在隐藏目标流量的静态指纹特征和动态指纹特征.本节通过分析拟态混淆技术总结其技术框架,分析常见实例.

1.2.1 拟态混淆框架

拟态混淆技术框架包括一条拟态管道和两个端点(一个是拟态客户端,另一个是拟态服务端).拟态客户端负责报文加密、整形,拟态服务端负责恢复、解密.拟态混淆技术可以形式化地描述为

$$P' = Shape(P,S,D),\ P = Shape^{-1}(P',S,D),$$

其中,$Shape()$是整形操作,$Shape^{-1}()$是整形逆操作,S 表示源报文协议参数,D 为目标协议参数.

图 3 所示为客户端发送报文 P.拟态客户端与拟态服务端以某普通协议建立连接,P 经拟态客户端整形为类似协议的 P',P'经审查网络到达拟态服务端,还原为 P 后发往服务端.审查者视 P'为正常报文.通常,拟态客户端和拟态服务端作为客户端和服务端集成组件.

Fig.2 Architecture of mimicry obfuscation technology

图 2 拟态混淆技术框架

1.2.2 典型实例

Kevin 提出 FTE[24]拟态混淆技术.FTE 混淆技术的基础是 FTE 模块:将密文正则表达式作为输入,指定正则表达式密文作为输出的整形模块.FTE 预先建立密钥集并引入缓存、编码、解析和解码 FTE 信息的 Record 层.FTE 混淆技术将待发送的报文模拟成普通报文.虽然流量经过 FTE 整形后具有较好的隐蔽性,但内容长度字段不匹配.

CensorSpoofer[25]是基于 IP 欺骗和模拟会话发起协议(SIP)VoIP 语音通信的混淆技术.CensorSpoofer 代理用 dummy 主机 IP 响应.代理按照 SIP、RTP/RTCP 报文模式发送报文,填充报文调整报文长度分布使其看似来自 dummy 主机的 VoIP 流量.

Mohajeri 提出的 SkypeMorph[26,27]分为 Setup 阶段和流量整形阶段.在 Setup 阶段,客户端通过带外信道获取 SkypeMorph 拟态客户端 Skype ID,并据此接入 Skype 网络:利用 Skype 通信交换密钥并协商与 SkypeMorph 客户端的通信端口,向 SkypeMorph 拟态客户端发起视频通话请求,SkypeMorph 拟态客户端检测到请求后忽略,然

后在协商端口上侦听客户端数据流.在流量整形阶段,oracle 模块提供 naïve 方法和 Traffic Morphing 方法,其中,Naïve 方法学习 Tor、Skype 流量的报文大小和间隔 n 阶分布以及预计算的模拟矩阵,Traffic Morphing 方法利用 morpher 库计算期望报文大小,利用 packetizer 调整报文时延.

1.3 隧　道

定义 5(隧道流量混淆). 将目标流量报文封装进正常流量报文的加密负载中,使目标流量难以从观测流量集识别的状态称为隧道流量混淆.

用样本流量隧道传输目标流量的技术称为隧道流量混淆技术,被认为是拟态混淆技术进阶.经过对隧道技术典型案例的分析,本节总结混淆技术通用框架,介绍常见实例并分析隧道混淆技术的隐蔽性.

1.3.1 隧道混淆技术框架

隧道混淆技术利用普通报文封装并传输非正常报文,经代理将非正常报文迭代转发到目的服务端.可形式化地表述为

$$P' = Header + Channel(E(P) + F),$$

$$P = D(Channel^{-1}(P' - Header), F),$$

其中,$Header$ 表示普通报文头部字段,$Channel()$为隧道运算过程,$Channel^{-1}()$表示隧道逆运算过程.

隧道混淆技术框架如图 6 所示:客户端在发送报文 P 之前预先与代理建立一条隧道(如 TLS),利用隧道将 P 转发到代理,代理将 P 发送到服务端.审查者仅看到普通报文.

Fig.3 Architecture of tunneling obfuscation technology

图 3 隧道混淆技术框架

1.3.2 典型实例

Brubaker 提出 CloudTransport[28].CloudTransport 用户需要安装 CloudTransport 客户端并注册代理云服务(如 Amazon S3)会合帐户(rendezvous account).用户选择 CloudTransport 网桥,用引导协议将会合帐户访问凭据(credentials)发至网桥.CloudTransport 客户端用云存储服务的标准库将报文传至会合帐户,网桥收集并转发到目的地.CloudTransport 仅对云服务代理前的流量起混淆作用,经 CloudTransport 网桥后,流量混淆作用被剥去.

Meek[29]被认为是目前最有效的混淆技术之一.用户使用 Tor 浏览器访问受审查网站.Tor 浏览器发送报文前被 Meek 客户端重新封装:Meek 客户端使用域名前置技术将受审查 URL 置于 TLS 加密的 HTTP Host Header 字段,前置 SNI 被设置成非审查 allow.example.审查者只看到 SNI.报文经 allow.example 服务器解析出 forbidden.example,并据此转发到运行 Meek-server 的 Tor 网桥.与此类似,Telex 利用审查机构非持续阻断的猫鼠游戏管控策略,将非阻塞网站作为代理[40].

Karlin 等人提出的诱骗路由(decoy-routing)技术[31,41]又称为折射网络技术.诱骗路由无需客户端连接拥有静态地址的代理,而发送 TCP 请求到诱骗目标主机(decoy dest),此 TCP 连接路径中存在诱骗路由器(decoy router),诱骗路由器将请求发送到诱骗代理(decoy proxy),代理与真实目的主机(convert dest)通信,将报文传回客户端.

1.4 混淆技术对比分析

为评估混淆技术,本文引入评价指标,包括:隐蔽性、计算开销和部署难度.

(1) 隐蔽性

隐蔽性是混淆流量抵御观测者识别的能力.谭庆丰等人[33]提出用相对熵描述混淆流量特征分布与普通流量特征分布的偏差,并据此提出隐蔽性度量方法.

相对熵:

$$D_s[p_\pi,q_\tau]=\sum_{o_{n,i}\in O_i}p_\pi(o_{n,i}\mid s)\ln\frac{p_\pi(o_{n,i}\mid s)}{q_\tau(o_{n,i}\mid s)}\ \text{iff}\ p_\pi=q_\tau, D[p_\pi,q_\tau]=0,$$

$$\widehat{D}=\frac{1}{|S|}\sum_{i=1}^{S}D_s[p_\pi,q_\tau]-D_s[p_\pi,q_\tau],$$

其中,p_x 为匿名通信系统第 i 维特征向量 O_i 的概率分布,q_τ 为目标协议在 O_i 上的概率分布,其中,S 为观测到的报文外显行为特征集,s 为匿名通信节点状态.

最大熵:

$$D_M=\max_{s\in S}D_s[p_\pi,q_\tau].$$

本文根据谭庆丰[33]提出的匿名系统隐蔽性度量,结合混淆流量识别方法提出隐蔽性度量:

$$d=\sum_{n=1}^{N}\frac{\widehat{D_n}}{D_{M_n}},$$

其中,d 是匿名通信系统不可观测性的量化,d 越小,隐蔽性越好,N 表示用于混淆流量识别的特征数量.

随机化混淆技术对流量特征字段、动态特征进行随机化运算,但可依据首个报文 URI 的相对熵,协议 π 与协议 τ 有明显差异,其隐蔽性度量为

$$d\approx 1-p_\tau\ln(p_\tau/q_\tau)/p_\pi\ln(p_\pi/q_\tau).$$

隐蔽性度量值较小.拟态混淆技术在加密的基础上按样本流量特征调整目标流量特征,包括报文格式、特征字段及报文间隔这 3 个特征,协议 π 与协议 τ 有明显差异,隐蔽性度量为

$$d\approx n\cdot[1-p_\tau\ln(p_\tau/q_\tau)/p_\pi\ln(p_\pi/q_\tau)].$$

拟态混淆技术较随机化混淆技术有更多的识别特征,隐蔽性较差.隧道混淆技术将目标流量报文加密封装进普通流量报文,协议 π 与协议 τ 相同,其隐蔽性度量值为

$$d=2\cdot[1-p_\tau\ln(p_\tau/q_\tau)/p_\pi\ln(p_\pi/q_\tau)].$$

随着报文填充、时延调整等方法的引入,

$$d<1-p_\tau\ln(p_\tau/q_\tau)/p_\pi\ln(p_\pi/q_\tau),$$

隐蔽性得到增强,隐蔽性最好.

(2) 计算开销

计算开销是流量混淆技术在流量转发过程中混淆运算的资源消耗量,涉及运算时间、运算次数、资源等.

计算开销评估形式化地描述为

$$P=\sum_{t=1}^{N}T_t\cdot n_t\cdot\sum_{i=1}^{I}S_i\cdot m_i,$$

其中,S_i 表示第 i 种资源数量(如 S_1 表示内存,S_2 表示 CPU),m_i 表示第 i 种资源调度次数,T_t 表示第 t 种混淆运算所需时间(如 T_1 表示加密时间,T_2 表示解密时间),n_t 表示第 t 种混淆运算的次数.

随机化只需两端调制解调处理,计算开销 $P=\sum_{t=1}^{2}T_t\sum_{i=1}^{2}S_i$,操作简单、速度快;拟态技术除加解密外,还引入两端格式调整,增加了两次调整开销,计算开销 $P=\sum_{t=1}^{4}T_t\sum_{i=1}^{2}S_i$;隧道技术将非正常流量加密封装进正常流量中,但代理对非正常流量的出隧道处理增加了一次开销,计算开销 $P=\sum_{t=1}^{3}T_t\sum_{i=1}^{2}S_i$.因此,随机化技术性能优于隧道技术,隧道技术性能优于拟态技术.

(3) 部署难度

流量混淆技术部署过程的难易程度是影响用户体验的重要因素.随机化混淆技术在发送端和接收端集成报文调制解调模块,安装客户端软件即可完成部署,操作简单.拟态混淆技术需要在互联网中部署多个模拟客户端和模拟服务端用于流量整形和还原,用户 PC 需要连接模拟客户端后才可以通信,部署难度较随机化混淆技术要大.隧道混淆技术将隧道封装模块集成在 PC 客户端,网络中部署一个代理节点,相对拟态混淆技术而言部署

简单,但比随机化混淆技术部署难度大.

在对比上述 3 个指标的同时,表 1 也从实现语言、应用场景、密钥交换和性能等角度对混淆技术进行了对比.其中,流量混淆技术应用场景主要包含匿名通信网或虚拟专用网.可以看出,在匿名通信网场景中,以保护用户隐私为首要目标,规避审查能力次之,故可选混淆技术范围较广,以 Tor 为代表的匿名通信网中使用的流量混淆技术种类也较多;而在虚拟专用网场景中,更看重规避审查能力和性能,故可选流量混淆技术较少,以 Lantern、Psiphon 为代表的虚拟专用网所使用的混淆技术有限,主要采用 Obfs 和域名前置技术.

Table 1 Flow obfuscation technology summary table

表 1 流量混淆技术汇总表

流量混淆技术		语言	应用场景		密钥交换	性能	隐蔽性	计算开销	部署难度
类型	具体技术		匿名通信网	虚拟专用网					
随机化	Obfs4	Go	Tor	Lantern	带内	高	中	$\sum_{t=1}^{2}T_t\sum_{i=1}^{2}S_i$	简单
	Obfs{2,3}	Python	Tor	–	带外	高			
	Basket2	Go	Tor	–	带内				
	ScrambleSuit	Python	Tor	–	带内	高			
	Bananaphone	Python	Tor	–	带内				
	LODP	C	Tor	–	–				
	Dust2	Haskell	Tor	–	–				
	Hexchat	Python	Tor	–	–				
	Git	Python	Tor	–	–	低			
	sshproxy	C	Tor	–	–				
	lampshade	Go	Tor	–	–				
	Dust	Python	Tor	uProxy	带外	高			
拟态	FTE	Python、C++	Tor	–	带外	高	低	$\sum_{t=1}^{4}T_t\sum_{i=1}^{2}S_i$	复杂
	StegoTorus	C++	Tor	–	带内	低			
	Marionette	C++	Tor	–	带外	中			
	Castle	C++	Tor	–	–	–			
	Code Talker Tunnel	C/C++	Tor	–	带外	低			
	SkypeMorph	C++	Tor	–	带外	低			
隧道	Cirripede	Assembly	–	–	带内	高	高	$\sum_{t=1}^{3}T_t\sum_{i=1}^{2}S_i$	一般
	CloudTransport		CloudTransport	–	带内	中			
	域名前置	Go(i.e.Meek)	Tor	Psiphon,Lantern, uProxy,GoAgent	带内	中			
	Telex	Python	–	–	带内	高			
	Non-PT Tor		Tor	–	带内	高			
	Decoy Routing	C	Decoy Routing	–	带外	高			
	Flashproxy	Python,Go, JavaScript	Tor	–	–	中			
	SnowFlake	Go,C/C++	Tor	–	–				

2 混淆流量识别技术

混淆技术虽然将非正常流量隐藏于普通流量,但审查者仍可基于微小差异研究新流量识别技术.混淆流量隐藏了原有流量的特点,如端口、IP 地址等.传统流量识别技术已无法有效识别混淆流量.为了有效识别混淆流量,研究新的流量识别方法势在必行.本文依据特征提取方法和实现原理将混淆流量识别技术分为两类:基于深度包检测[42]和基于机器学习的混淆流量识别技术.

2.1 基于深度包检测的流量识别技术

2.1.1 面向随机化的流量识别技术

(1) 基于熵的流量识别

基于报文熵值的分析最早用于网络异常检测[43]和 P2P[44]、VoIP[45]流量分类.随机化混淆技术的出现使基于特征字符的流量识别技术应用大幅减小.研究人员开始考虑将熵应用到混淆流量的检测和监控中[45].传统的加密协议(如 TLS)报文的握手信息包含未经加密的固定字符串集合.随机化混淆技术 Obfs 对每条流所有报文

加密,使每条流第 1 个报文的熵值可作为可靠的识别依据.Wang[31]利用 Obfs{3,4}前 2 048 字节的熵识别经报文长度过滤的流量,获得 100%的识别率.为了便于熵值计算和识别,Wang[31]设计了 Shannon-entropy 仿真器.基于仿真器计算传统的 KS(Kolmogorov-Smirnov)双样本测试可以有效地提高识别效率.Brandon[46]分析 Dust、Obfs、SSL 流量,建立贝叶斯预测模型,利用第 1 条报文熵识别混淆流量,其准确度高达 94%.

2.1.2 面向拟态的流量识别技术

(1) 基于特征字符的流量识别

Wang 等人[31]发现,StegoTorus 在传输 PDF 文件过程中可依靠检查 PDF 文件的 xref 表识别文件传输(尤其是 PDF 文件),并为此划分为 4 类:标准(standard)、异常(malformed)、部分(partial)和其他(other),分别依据“%PDF%”“%%EOF”、xref 关键字和状态代码“206 Partial Content”识别.Houmansadr 等人[47]利用 StegoTorus 的 HTTP 响应特征识别并根据请求类型将响应分为 6 类:GET long、GET non-existing、HEAD existing、OPTIONS common、DELETE existing、TEST method、GET Wrong Protocol,分析了误报的原因和概率.Wang 等人[31]通过实验发现,FTE 报文虽然经过加密转换,但 HTTP 的 Content-Length 字段与真实内容长度不匹配,据此获得的误报率低于 4%.基于特征字符的流量识别技术开销相对较小,有很高的准确率,但前期特征发现和提取过程工作量较大.

(2) 基于熵的流量识别

FTE 混淆流量经过变换处理看似普通流量,但第 1 个 HTTP GET 报文中 URI 经加密看似随机字节.经 Wang 等人[31]实验发现,FTE 的 URI 熵落在 5.5~5.8 比特相对窄的范围内,非 FTE 的 URI 熵均小于 5.1 比特.据此熵识别 FTE 流量,获得 100%的识别率.

2.1.3 面向隧道的流量识别技术

(1) 基于协议字段的流量识别

何高峰[18]提出利用 TLS 握手协议加密套件、数字证书序列号与普通流量的差异识别 Tor 流量的方法,识别率为 100%.何高峰等人总结出了 7 个稳定的 Meek 流特征:单一链接特征、有序连接特征、TLS Cipher Suits 特征、TLS Extensions 特征、TLS Server Name 特征、轮询请求特征、分组传输特征,并利用其中的静态指纹特征过滤 Meek 流量[32].

(2) 基于报文长度的流量识别

基于 Tor 的混淆流量中报文长度与信元长度(512 字节)存在倍数关系.何高峰[18]按照信元及发送策略分析混淆流量报文长度分布,将报文长度按出现频率由高到低排序并求频率之和,选取大于门限值频率的长度作为特征长度.统计 Tor 上行流量和其他类型流量中相同特征报文出现的频率形成长度分布,其离线流量识别率达到 95%,在线流量识别率达到 91%,误报率只有 1.2×10^{-5},填充技术的引入使基于报文长度分布的识别技术不再适用.

(3) 基于熵的流量识别

何高峰[18]与何永忠[39]分别统计发送报文长度信息熵和接收报文长度信息熵,归一化预处理成{0,1}范围内的实数,利用 SVM 机器学习算法加以判别[18,39].吴震针对识别准确率较低的问题,提出一种基于信息熵的流量识别方法,用信息熵级联分簇,生成识别模型,识别率在 90%以上[48].谭庆丰[33]提出匿名通信系统的不可观测性度量方法,提出基于相对熵的混淆流量识别方法,从报文间隔分布相对熵和长度分布相对熵中发现普通 HTTPS 报文与 Meek 报文间存在明显的差别.

(4) 基于行为模式的流量识别

基于行为模式的流量识别技术,又称为启发式流量识别技术,通过匹配节点间通信模式推断节点关系或者特定角色[49].早期启发式识别技术利用 P2P 已知属性,如同时用 UDP、TCP 两种协议通信、利用独立连接传输大量数据,识别精度较低.Perenyi 等人扩大了参数匹配范围以提高识别精度,利用精确匹配降低误报[50,51].John 提出利用报文长度模式匹配识别基于 SSL 的 Tor 流量[52].何高峰通过提取目标 TLS 流量特殊长度报文,计算报文间隔,并将间隔序列带入轮询请求机制判断器进行判断,识别率为 97%[18,32].Sami 等人提出基于马尔可夫模型

的流量识别技术,通过分析虚电路构建过程并结合日志交叉分析,形成电路构建序列的马尔可夫模型(HMM),具有高达 98%的识别率[34].

据调查,深度包检测已被中国、伊朗、土耳其等国家用于网络审查[53].表 2 表明,国家层面使用的深度包检测技术主要基于静态指纹特征.

Table 2 Cases of country review based on depth packet detection

表 2 基于深度包检测的国家审查案例

国家	时间	技术	主动/被动
中国	2007 年	IP/DNS 黑名单,Tor TLS 头部指纹,Tor-relay 握手探针	主动+被动
伊朗	2012 年	IP 黑名单,关键字,TLS 握手指纹	被动
土耳其	2014 年	IP/DNS 黑名单,Twitter 和 Youtube 的 BGP 劫持	主动+被动
巴基斯坦	断续	IP/DNS 黑名单,URL,关键字,文件类型	被动
叙利亚	2011 年	IP/DNS 黑名单,即时消息指纹,社交媒体关键字黑名单	被动
埃塞俄比亚	2012 年	Skype 检测,Tor TLS 指纹	被动
白俄罗斯	2014 年	IP/DNS 黑名单,根据报文内容过滤	被动
哈萨克斯坦	2012 年	加密指纹,Tor TLS 握手指纹	被动
埃及	2014 年	IP/DNS 黑名单,即时消息指纹,社交媒体关键字黑名单	被动

2.2 基于机器学习的流量识别技术

随着人工智能的快速发展,越来越多的机器学习技术应用于流量识别,提高了流量识别速度和准确率.

2.2.1 机器学习算法在混淆流量识别中的应用

(1) 面向随机化混淆技术

Wang 等人[31]提取每个报文负载的最大、最小和平均熵特征,时间特征和报文头特征,为机器学习训练提出两种流量窗口策略:一条流的前 X 个报文或一条流的前 X 秒,选用分类算法(kNN、朴素贝叶斯或 CART)测试并识别混淆流量.Obfs3 的识别率为 97.2%;Obfs4 的识别率为 97%.

(2) 面向拟态混淆技术

Wang 等人[31]利用上面(1)中所述方法识别 FTE 流量,FTE 的识别率为 97.8%.

(3) 面向隧道混淆技术

何高峰等人[18]分析 Tor 流量典型报文长度并标记,按照寻找 SVM 最优分类超平面算法并获得 91%的识别率.Song 等人[54]定义并提取二元组$\{T,S\}$带宽特征(在时间 T 时,已传递 S 字节数据),用一条流前 8 个报文长度训练 SVM 分类器可将 Tor 流量与普通流量区别开.何高峰[18]与何永忠[39]二人通过归一化报文长度方差、长度熵等特征,利用 SVM 识别 Meek 流量.Song 等人[54]基于报文组建时大小不固定的特征,用 SVM 算法识别基于 TLS 或 Obfs 的 Tor 混淆流量.

为了获取更高的识别率和可用性,Alaeddin 等人[55]采集 Tor 流量和普通流量,提取每条流的总字节数、总报文数、流持续时间等 40 种流特征,分别对朴素贝叶斯、随机森林等算法训练并对普通 HTTP 流量和 Tor 流量进行分类.Wang[31]针对 Meek-Amazon、Meek-Google 两种隧道技术采用上述(1)中所述方法,获得 Meek-Amazon 识别率为 97.3%,Meek-Google 识别率为 98.3%.Shahbar 等人[56]从电路级和数据流级实现 Tor 流量应用的分类,包括 Browsing、Streaming、BitTorrent、电路级选取发送的信元数、上行流量信元总数、下行信元与上行信元比等特征,数据流级采用 Tranalyzer2、Tcptrace 等软件自动生成特征,分别采用贝叶斯网络、朴素贝叶斯、C4.5、随机森林等算法加以训练和识别.LashKari 等人[57]提取基于时间的 32 种流特征,包括流持续时间、带宽、上行/下行报文间隔等,用 kNN、C4.5 决策树算法识别 Tor 流量,准确率达到 92%.Deng 等人提取源端口、目的端口、总报文数等 35 种数据流特征,将每一条流视为一个粒子并定义粒子间操作,用重力聚类算法解决 Tor 流量分类问题,利用 ExperimenTor 环境收集流量并加以测试,测试效果优于 DBSCAN、K-means 等聚类算法[58].Hodo 等人[59]在 LashKari 的研究基础上用 ANN、SVM 算法对 Tor 流量进行识别,获得了 95%的准确率.Lotfollahi 等人[60]提取 IP 头部 20 字节、TCP/UDP 头部 20 字节(UDP 填充 12 字节零)和负载 1 460 字节作为输入,用 CNN、SAE 深度神经网络对 Tor、网页、语音、视频等 17 类流量进行识别,达到 95%的精确率.深度学习方法减少了

传统机器学习提取流特征的开销,但特征数明显增多,造成巨大的训练开销.虽然深度学习方法取得了良好效果,但数据集规模小、扩展性差,真实的大规模环境中效果有待验证.LashKari、Hodo、Lotfollahi 等人的实验均使用纽布伦斯威克大学公开数据集[61].

2.3 混淆流量识别技术对比分析

流量混淆技术旨在规避审查,保护用户隐私.混淆流量识别技术是识别混淆流量、获取流量信息的网络攻击技术.攻击混淆流量伴随着审查规避系统的不断加强,混淆流量识别技术相应地推陈出新.两种技术互为攻防,发展过程中此消彼长.

基于 DPI 的混淆流量识别技术从 SSL 报文或应用层报文中获取特征字段、统计特征,但需要人工发现并提取流量特征等大量预处理工作.基于特征字段的 DPI 可依靠简单的字段匹配,复杂度为常数 C.随着新的流量混淆技术引入,DPI 流量识别方法依靠统计特征识别流量变得愈发困难,准确率开始降低.DPI 流量识别技术只对当前流量识别有较好的识别能力,对未来出现的混淆流量需重新分析和提取,扩展性差.

现已应用于流量识别的机器学习算法包括 4 类:有监督学习、无监督学习、半监督学习和集成学习[62].多种机器学习分类器协同处理可以有效提高识别精确度[63,64].基于机器学习的混淆流量识别技术是为了提高流量识别率、减少人类工作量提出来的.浅层机器学习技术仍依靠人工提取流量特征,其广泛采用 BP 算法,尤其是 SVM(复杂度)、决策树(复杂度)等算法;深层机器学习(深度学习)技术主要依靠流量的 Bit 串输入识别流量,无需人为提取特征,节省了大量人力、财力和物力,广泛采用 CNN(复杂度)、SAE(复杂度)等深度神经网络算法.机器学习技术可用于现有及未来所有流量的识别,具有很强的扩展性.

流量识别技术对比情况见表 3.针对随机化混淆流量的识别特征单一,主要依靠熵,但识别率高;拟态流量识别特征包括特征字符和熵;隧道技术识别方法丰富,包括协议字段、报文长度、熵和行为模式,除基于协议字段的识别方法可达到 100%的识别率以外,其他方法的识别率为 97%左右.机器学习可结合多个特征,经训练可获得较高的识别率,但需人工发现并提取特征;深度学习可在没有任何特征时通过训练获得更高的识别率,但训练开销大,周期长.从表 3 可以看出,由于 DPI 依靠明显的流量特征,识别率高;浅层机器学习需要数十个流量特征,确定和提取流量特征依然需要大量工作,深度学习技术无需人为提取特征,较浅层机器学习识别方法可节省大量工作,具有更高的识别率.今后深度学习流量识别技术将是流量识别技术的未来发展趋势.

Table 3 Flow identification technology comparison

表 3 流量识别技术对比表

混淆流量识别技术			混淆技术	识别特征	复杂度	识别率(%)
DPI	随机化	熵	Dust,Obfs	报文熵	$O(n\log n)$	100
	拟态	特征字符	StegoTorus, FTE	特征字符	C	
		熵	FTE	URI 熵	$O(n\log n)$	100
	隧道	协议字段	Tor,Meek	证书序列号,加密套件,TLS 握手协议扩展字段	C	100
		报文长度	Tor	特征报文长度	$O(n)$	95
		熵	Meek	长度熵	$O(n\log n)$	97
		行为模式	Meek,Tor	长度、时间序列、虚电路建立状态	$O(n)$	97
机器学习	随机化	*k*NN、朴素贝叶斯或者 CART	Tor	报文负载的最大、最小和平均熵特征,时间特征和报文头特征	$O(m^2n^2)$	97.2
	拟态	*k*NN、朴素贝叶斯或者 CART	Tor	报文负载的最大、最小和平均熵特征,时间特征和报文头特征	$O(m^2n^2)$	97.8
	隧道	SVM	Tor,Meek	长度分布、长度方差、长度熵分布等 7 种特征	$O(n)$	91
		朴素贝叶斯,*k*NN,C4.5,随机森林等决策树	Tor	流持续时间,传输速率,上行和下行报文间隔等 32 种特征	$O(m^2n^2)$	92
		聚类算法	Tor	端口、总报文数等 35 种数据流特征	$O(n^3)$	92
		深度神经网络(CNN,SAE)	Tor	IP 头部,TCP/UDP 头部及报文负荷特征	$O(n^3)$	95

3 混淆流量追踪技术

流量追踪技术是观察者确定流量发送者和接收者之间通信联系的技术.根据追踪技术对流量是否干涉可分为被动关联技术和主动关联技术.本节将按被动关联技术和主动关联技术两类深入介绍典型流量追踪技术.

3.1 被动关联技术

被动关联技术是依靠分析被动观测的流量特征关联发送者和接收者通信联系的技术.观测者不对流量做任何干扰.吕博等人[64]总结了关于被动关联技术的相关研究,本文在此基础上进一步总结并将其重新划分为揭露分析攻击、流量分析攻击和指纹攻击.

3.1.1 揭露分析攻击

Berthold 提出交集分析攻击[65],利用相对较小的通信对象集进行分析,利用不同消息在网络中使用同样的路径进行关联分析.Kesdogan 等人在交集分析的基础上提出揭露分析[66],假设用户使用混淆流量和固定大小的用户集合通信,长期观察特定用户发送消息时的接收用户集,通过并集关联通信对象.为了提高分析效率,Danezis 等人在揭露分析的基础上提出统计揭露方法[67],分别收集用户发送消息与不发送消息时接收者的统计特征来关联发送者与接收者.Qin 提出的 STARS[68]利用原生流量统计特征分析发现端到端的通信关系.STARS 的流量关联方法虽然可以有效地关联匿名流量,但需经端到端流量矩阵和概率分布等复杂运算,开销较大.Mallesh 等人针对统计揭露分析攻击技术提出 cover traffic 模拟用户发送模式,虽然不能完全抵御统计揭露攻击,但可消耗攻击者更长的时间[69].Rajiv 提出的基于权重的关联方法可获得更高的精确度[70].Herrmann 等人揭露密钥信息,达到关联流量的目的[71].

3.1.2 流量分析攻击

流量分析攻击涵盖范围广,通常包括消息编码攻击(message coding attack)、时间攻击(timing attack)、通信模式攻击(communication pattern attack)、交叉攻击(intersection attack)、报文容量攻击(packet volume attack)、报文计数攻击(packet counting attack)流量分析攻击等攻击方法.Murdoch 等人[72]通过部署 Tor 节点探针获取 Tor 节点流量,分析其报文大小及延时特征,利用关联函数完成追踪目的.Volker 等人[73]提出时隙报文计数方法,通过计算关联系数期望值和偏差来关联 Tor 流量.Song 等人[54]在识别基于 TLS 或基于 Obfs 的 Tor 流量的基础上提取{时间,流大小},利用 *k*-means 算法分簇、匹配入口节点和出口节点的流量,以便关联追踪.

3.1.3 指纹攻击

指纹攻击(fingerprinting attack)是基于通信协议特征的追踪方式.Steven 提出利用主机时钟倾斜(高达 50ppm)作为指纹揭露隐藏服务的追踪方式[74].虽然 Weinberg 提出的 StegoTorus 混淆技术提高了 Tor 抵抗指纹攻击的能力[75],但 Biryukov 等人通过测量隐藏服务的访问量来解密隐藏服务,利用指纹环(fingerprint circle)信息绕过混淆技术,达到追踪的目的[76].Liberatore 等人根据数据包长度序列指纹特征,利用朴素贝叶斯分类器追踪 HTTP 报文[77].Wang 等人提出将 *k*NN 分类器应用于带权重的大量指纹特征数据集,以识别用户网络活动并获得比 Liberatore 更高的精确度[78].Kwon 等人利用 Tor 网络 Circuit 建立过程与普通链路的差异提出 circuit 指纹攻击,将精度提高到 99%[79].Hayes 等人提出基于 *k*NN 的 *k*-fingerprinting 攻击方法,提高了追踪精度,但是应用范围局限于暗网网站的特定网页[80].Zhuo 等人提出基于配置文件的隐马尔可夫模型(PHMM)的网站建模方法,可用于追踪 SSH 和 shadowsocks 等混淆流量[81].Juarez 通过有监督分类器分类用户访问的网页,利用网站活动指纹有效地攻击 Tor 网络[82].虽然近几年使用机器学习,提高了基于指纹攻击的追踪技术准确率,但是追踪方法趋于单一,局限于追踪网页流量.

3.2 主动关联技术

主动流量关联技术是攻击者对目标流量采取主动干涉手段将接收端流量与发送端流量关联起来的技术.主动关联技术操作简单,开销低,且实时性强,适用于任何流量.

3.2.1 流水印

研究人员提出主动网络流水印(active network flow watermark,简称 ANFW)这一概念[83].如图 14 所示,为了

证实爱丽丝与鲍勃之间存在通信关系,在爱丽丝出口处设置调制器将水印信息嵌入流量,然后在鲍勃接收前设置检测器检测水印信息,如果水印信息匹配,则证明爱丽丝与鲍勃之间具有通信关系.ANFW 根据水印嵌入方式,将流水印分为基于流速的流水印和基于时间特征的流水印.

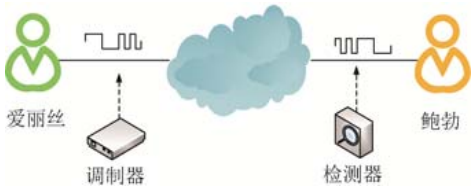


Fig.4 Architecture of flow watermark technology

图 4 流水印使用方法

(1) 基于流速

基于流速的流水印技术主要依靠调制流量发送速率.扩频是调制流量发送速率的典型方法.在物理层对发送信号按照某种扩频函数(如利用伪噪声 pseudo-noise,简称 PN)扩展频带宽度.扩频函数就是水印嵌入方法,嵌入的信息被称为水印信号.直序扩频(DSSS)[35]是扩频水印的典型代表.追踪者对原始信号 D_s 加入水印(PN_s 码)信号后得到信号 S_s,经路由转发(假设未受干扰)后,追踪者提取信号 S_r,如果 S_S=S_r,则利用 PN_r 逆运算可恢复原始信号 D_r.

$$D_r=\frac{\sum S_r\cdot PN_r}{N}=D_s\frac{\sum PN_s\cdot PN_r}{N}.$$

扩频流水印提供一个隐蔽、实时的流量追踪技术.目前尚未有允许用户消除扩频流水印的解决方案.

(2) 基于时间

基于时间的流水印技术分为两种:(1) 基于报文间隔;(2) 基于时隙分割.基于报文间隔的流水印技术通过调整间隔嵌入水印.Wang 在 2003 年提出的 IBW 方法通过随机选取流内两个包分组,调整分组到达或离开的时间间隔以实现水印注入[84].为了解决 MFA 攻击威胁,Houmansadr 提出 SWIRL[85].SWIRL 算法虽然具有良好的多流攻击、拥塞攻击抵御能力,但易受抖动和垃圾包注入的干扰,鲁棒性较差.基于时隙分割的流水印技术按照时隙分组嵌入水印.基于时隙分割流水印技术的典型实例是基于时隙质心的流水印技术[36].将 2n 个时隙按照水印信号的 bit 数分为 2 个组,每个组包含 L 个小组,每个小组对应 n/L 个时隙.如下计算各小组时隙质心:

$$Cent(I_i)=\frac{1}{n_i}\sum_{j=0}^{n_i-1}\Delta t_{ij}.$$

计算两个群中对应同一水印 bit 的时隙差.水印调制模块根据差值决定每个组的延时增量.

基于时隙质心的流水印技术虽然具有较好的隐蔽性和抗干扰能力,但 Kiyavash 针对 IBW、ICBW 提出多流攻击(MFA)[86].Luo 等人将 ICBW 与 DSSS 相结合,提出基于直序扩频的时隙质心流水印方法(interval centroid based spread spectrum watermarking,简称 ICBSSW)[87],在应对 MSAC 攻击和 MFA 方面具有较好的效果.同时,具有追踪多条流的能力,但算法复杂度高,开销大,实用性低.Wang 提出的 DICBW[88]在抵御 MFA、网络干扰、流分割与合并等方面有较好表现.

3.2.2 渗 透

(1) 中间人

混淆技术难以抵御中间人攻击.审查者提出基于 HTTP 的中间人攻击,利用受控节点嵌入指定数量图片标签的页面,发现客户端与 Web 服务器通信[89].嵌入图片增加了通信开销,隐蔽性差.研究者利用受控出口节点在 HTTP 中嵌入 JavaScript 或 HTML 代码,进行中间人攻击[90].基于 botnet 的技术,利用 bot master 控制大量沦陷的网络节点监控网络活动[91].卡内基梅隆大学研究员 Michael 和 Alexander 提出了打入受控卧底节点破解 Tor 网络的方法,这与 FBI 侦破丝绸之路的方法不谋而合.为了提高追踪效率,Murdoch 和 Danezis 提出 Circuit Clogging 方案,用探针探测 Tor Relay 节点流量并假冒服务器做出回应[92].

(2) 节点发现

VPN 只有一个代理节点.利用混淆流量识别技术即可发现 VPN 代理节点.但 Tor 中继节点信息或网桥信息是非公开且变化的,混淆技术的引入增强了 Tor 中继节点和网桥的隐蔽性,基于 Tor 的混淆流量追踪具有很大的挑战性.Mclachlan 等人提出基于大量邮件和 HTTP 服务器中包含的隐藏网桥信息进行枚举攻击[93].Winter 和 Ensafi 等人推断 GFW(The Great FireWall of China)通过流量识别技术和节点发现攻击技术确认发往 Tor 网桥的混淆流量,并调度扫描节点伪造连接请求以尝试连接 Tor 网桥[94,95].

(3) 重放

重放攻击重复发送通信中被截取的报文,干扰信息的正常接收.假设攻击者控制某节点复制混淆流量,沿相同方向再次发送相同报文就会扰乱 Tor 节点计数器计数,造成解密失败[47].通过受控恶意入口节点复制、篡改发送的报文导致出口节点无法识别[96].Zhen 提出基于 Tor 的发现、阻断和追踪恶意流量的系统 TorWard[97].TorWard 在 Tor 出口节点部署入侵检测系统(IDS),用于 Tor 恶意流量的检测、阻断和追踪.TorWard 中出口节点作为代理提取转出流量信息,交给自动管理工具后重新将流量注入 Tor 网络中发往服务端.

流量追踪技术对比分析

本节汇总流量追踪技术,细节可见表 4.从汇总表可以看出,被动关联技术包括揭露分析、流量形状和流量指纹技术.但是 3 种方法均需在网络中部署探针被动采集大量流量,并做大量分析计算工作,实时性差.Song 等人使用 *K*-means 聚类算法实现 Tor 入口流量和 Tor 出口流量的关联,为被动关联技术提高追踪效率提供借鉴[54].主动关联技术以流水印技术和渗透技术为主.两种主动关联技术都可以简单、有效地达到追踪目的,但是流水印技术容易受到报文重放、篡改、乱序等情况的干扰,渗透技术部署难度大、成本高.

被动关联技术以流量识别技术为基础,对流量特征依赖性较强,故对随机化流量、拟态流量追踪能力较差.主动关联技术中的流水印技术操作简单,精度高,可以追踪任何混淆流量,因此将会是未来发展的趋势.渗透技术因可有效探知任何混淆流量,可同时追踪多种流量,自产生至今一直沿用.但其部署难度大、成本高是影响其广泛使用的重要因素.如何克服这些弊端将是研究渗透技术的未来研究重点.

Table 4 Flow tracking technology summary

表 4 流量追踪技术汇总表

分类	追踪技术		追踪障碍	混淆工具	精确率
被动关联技术	揭露分析技术		交集范围限制	任何混淆技术	与流量采集规模有关
	流量形状攻击		重放、广播等	任何混淆技术	与流量采集规模有关
	流量指纹技术		混淆能力	任何混淆技术	与流量采集规模有关
主动关联技术	流水印技术	基于流速	多流攻击	任何混淆技术	–
		基于时间	丢包、重组、乱序、多流攻击	任何混淆技术	与时隙内报文个数相关
	渗透技术	节点发现	节点信息隐藏	Tor	追踪到第 1 跳节点前
		中间人	成本、部署	任何混淆技术	与受控节点数量有关
		重放攻击	时间检测	任何混淆技术	–

4 总 结

本文从当前审查规避系统的背景入手,描述了流量混淆技术的重要性,分析了当前比较重要的 3 类流量混淆技术,总结了混淆技术框架并分析其隐蔽性.从混淆技术出发,进一步探讨了针对混淆流量的识别技术,并将其按照混淆技术类型分为基于深度包检测的流量识别技术和基于机器学习的流量识别技术.随着网络的发展和人工智能的广泛应用,实时性和智能化将会成为流量识别的趋势.为了进一步威慑非法网络行为,审查者开始研究流量追踪技术.流量追踪技术包含被动关联和主动关联技术两种.被动关联存在开销大、周期长等弊端,机器学习技术在流量分析上具有高效、准确等特点,将是未来研究的方向.主动关联技术减少了数据处理规模和计算开销,但流水印技术抗干扰能力差,难以抵抗多流攻击等,而渗透技术部署难度大、成本高.流水印技术的当务之急是提高抗干扰能力和抵抗攻击能力,而轻量型低成本是渗透技术未来的研究方向.

References:

[1] Chaum DL. Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM, 1981,24(2): 84–90. [doi: 10.1145/358549.358563]

[2] Boyan J. The Anonymizer: Protecting user privacy on the Web. Computer-Mediated Communication (CMC) Magazine, 1997.

[3] Reiter MK, Rubin A D. Crowds: Anonymity for Web transactions. ACM Trans. on Information and System Security (TISSEC), 1998,1(1):66–92. [doi: 10.1145/290163.290168]

[4] Hao F, Zielinski P. A 2-round anonymous veto protocol. In: Proc. of the Security Protocols Workshop. 2006,5087:202–211. [doi: 10.1007/978-3-642-04904-028]

[5] Sherwood R, Bhattacharjee B, Srinivasan A. P5: A protocol for scalable anonymous communication. Journal of Computer Security, 2005,13(6):839–876. [doi: 10.1109/SECPRI.2002.1004362]

[6] Zantout B, Haraty R. I2P data communication system. In: Proc. of the ICN. 2011. 401–409.

[7] lantern. https://github.com/getlantern/lantern

[8] Kean S. Internet research, uncensored. Chronicle of Higher Education, 2007,53(29).

[9] SnapVPN. https://snap-vpn.cn.uptodown.com/android

[10] jisuVPN. http://jisuvpn.msnyou.com/

[11] Pokemonvpn. https://play.google.com/store/apps/details?id= com.xxykj.pokemonvpn

[12] VPN Usage Around the World-Q2. 2017. Globalwebindex. https://cdn2.hubspot.net/hubfs/304927/Downloads/VPN-Usage-Around-the-World-Infographic.pdf

[13] China Officially Outlaws Unauthorised VPNs. 2017. https://advox.globalvoices.org/2017/01/23/china-officially-outlaws-unauthorised-vpns/

[14] Huang Y, Lin Y. Transnational cybercrime is getting worse and stronger, China and ASEAN seek to join hands. 2015 (in Chinese). http://www.12377.cn/txt/2015-09/15/content_8235494.htm

[15] Aliens C. Terrorist used Tor to connect with ISIS, source said. 2017. https://www.deepdotweb.com/2017/06/26/terrorist-used-tor-connect-isis-source-said/

[16] Censorship of Twitter. https://en.wikipedia.org/wiki/Censorship_of_Twitter

[17] Ministry of Industry and Information Technology of the People's Republic of China (in Chinese). http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757020/c5471946/content.html

[18] He GF, Yang M, Luo JZ, Zhang L. Online identification of tor anonymous communication traffic. Ruan Jian Xue Bao/Journal of Software, 2014,24(3):540–546 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4253.htm [doi: 10.3724/SP.J.1001.2013.04253]

[19] Wiley B. Dust: A blocking-resistant internet transport protocol. Technical Report, 2011. http://blanu.net/Dust.pdf

[20] Kadianakis G, Mathewson G. Obfs2 (the twobfuscator). 2011. https://gitweb.torproject.org/pluggabletransports/obfsproxy.git/tree/doc/obfs2/obfs2-protocolspec.txt

[21] Kadianakis G, Mathewson G. obfs3 (the threebfuscator). 2013. https://gitweb.torproject.org/pluggabletransports/obfs proxy.git/tree/doc/obfs3/obfs3-protocolspec.txt

[22] Angel Y ,Winter P. obfs4 (the obfourscator). 2014. https://gitweb.torproject.org/pluggable-transports/obfs4.git/tree/doc/obfs4-spec.txt

[23] Winter P, Pulls T, Fuss J. ScrambleSuit: A polymorphic network protocol to circumvent censorship. In: Proc. of the 12th ACM Workshop on Privacy in the Electronic Society. ACM, 2013. 213–224.

[24] Dyer KP, Coull SE, Ristenpart T, *et al*. Protocol misidentification made easy with format-transforming encryption. In: Proc. of the 2013 ACM SIGSAC Conf. on Computer & Communications Security. ACM, 2013. 61–72.

[25] Wang Q, Gong X, Nguyen GTK, *et al*. Censorspoofer: Asymmetric communication using IP spoofing for censorship-resistant Web browsing. In: Proc. of the 2012 ACM Conf. on Computer and Communications Security. ACM, 2012. 121–132.

[26] Mohajeri Moghaddam H, Li B, Derakhshani M, *et al*. Skypemorph: Protocol obfuscation for tor bridges. In: Proc. of the 2012 ACM Conf. on Computer and Communications Security. ACM, 2012. 97–108.

[27] Moghaddam MH. SkypeMorph: Protocol obfuscation for censorship resistance [MS. Thesis]. University of Waterloo, 2013.

[28] Brubaker C, Houmansadr A, Shmatikov V. Cloudtransport: Using cloud storage for censorship-resistant networking. In: Proc. of the Int'l Symp. on Privacy Enhancing Technologies Symp. Cham: Springer-Verlag, 2014. 1–20. [doi: 10.1007/978-3-319-08506-7_1]

[29] Fifield D, Lan C, Hynes R, *et al*. Blocking-Resistant communication through domain fronting. Proc. on Privacy Enhancing Technologies, 2015,2015(2):46–64.

[30] Karlin J, Ellard D, Jackson AW, *et al*. Decoy routing: Toward unblockable internet communication. In: Proc. of the USENIX Workshop on Free and Open Communications on the Internet. 2011. https://www.usenix.org/legacy/event/foci11/tech/final_files/Karlin.pdf

[31] Wang L, Dyer KP, Akella A, *et al*. Seeing through network-protocol obfuscation. In: Proc. of the ACM SIGSAC Conf. ACM, 2015. 57–69. [doi: 10.1145/2810103.2813715]

[32] Li X. Research and implementation of identification for Tor anonymous communication based on meek [MS. Thesis]. Beijing: Beijing Jiaotong University, 2016 (in Chinese with English abstract).

[33] Tan Q, Shi J, Fang B, *et al*. Towards measuring unobservability in anonymous communication systems. Journal of Computer Research and Development, 2015,52 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2015.20150562]

[34] Zhioua S. Tor traffic analysis using hidden Markov models. Security & Communication Networks, 2013,6(9):1075–1086. [doi: 10.1002/sec.669]

[35] Yu W, Fu X, Graham S, *et al*. DSSS-Based flow marking technique for invisible traceback. In: Proc. of the IEEE Symp. on Security and Privacy. 2007. 18–32. [doi: 10.1109/SP.2007.14]

[36] Wang X, Chen S, Jajodia S. Network flow watermarking attack on low-latency anonymous communication systems. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE, 2007. 116–130. [doi: 10.1109/SP.2007.30]

[37] Kiyavash N, Houmansadr A, Borisov N. Multi-Flow attacks against network flow watermarking schemes. In: Proc. of the USENIX Security Symp. 2008. 307–320.

[38] Jia W, Tso FP, Ling Z, *et al*. Blind detection of spread spectrum flow watermarks. Security and Communication Networks, 2013,6(3):257–274.

[39] He YZ, Chen M. Protocol mimicry technique and its development. Journal of Beijing Jiaotong University, 2016 (in Chinese with English abstract). [doi: 10.11860/j.issn.1673-0291.2016.05.001]

[40] Wustrow E, Wolchok S, Goldberg I, *et al*. Telex: Anticensorship in the network infrastructure. 2011. https://www.usenix.org/legacy/event/sec11/tech/full_papers/Wustrow.pdf [doi: 10.1.1.211.418]

[41] Karlin J, Ellard D, Jackson AW, *et al*. Decoy routing: Toward unblockable Internet communication. In: Proc. of the USENIX Workshop on Free and Open Communications on the Internet. 2011. https://www.usenix.org/legacy/event/foci11/tech/final_files/Karlin.pdf

[42] Wu Q. Design and implementation DPI and DFI-based system of flow identification and control [Ph.D. Thesis]. Chengdu: University of Electronic Science and Technology of China, 2013 (in Chinese with English abstract).

[43] Nychis G, Sekar V, Andersen DG, *et al*. An empirical evaluation of entropy-based traffic anomaly detection. In: Proc. of the Internet Measurement Conf. 2008. 151–156. [doi: 10.1145/1452520.1452539]

[44] Lu G, Zhang HL, Ye L. P2P traffic identification. Ruan Jian Xue Bao/Journal of Software, 2011,22(6):1281−1298 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3995.htm [doi: 10.3724/SP.J.1001.2011.03995]

[45] Wright CV, Ballard L, Monrose F, *et al*. Language identification of encrypted VoIP traffic: Alejandray Robert or Alice and Bob? In: Proc. of the Usenix Security Symp. 2007. 4.

[46] Wiley B. Blocking-Resistant protocol classification using Bayesian model selection. Technical Report, University of Texas at Austin, 2011.

[47] Houmansadr A, Brubaker C, Shmatikov V. The parrot is dead: Observing unobservable network communications. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE Computer Society, 2013. 65–79. [doi: 10.1109/SP.2013.14]

[48] Wu Z, Liu XB, Tong XM. Traffic identification method based on information entropy. Computer Engineering, 2009,35(20): 115–117 (in Chinese with English abstract).

[49] Barker J, Hannay P, Bolan C. Using traffic analysis to identify tor usage—A proposed study. In: Proc. of the 2010 Int'l Conf. on Security & Management, SAM 2010, Vol.2. 2010.

[50] Perényi M, Dang T D, Gefferth A, *et al*. Identification and analysis of peer-to-peer traffic. Journal of Communications, 2006,1(7): 36–46. [doi: 10.1109/ICIW.2010.36]

[51] John W, Tafvelin S. Heuristics to classify Internet backbone traffic based on connection patterns. In: Proc. of the Int'l Conf. on Information Networking, ICOIN 2008. IEEE, 2008. 1–5. [doi: 10.1109/ICOIN.2008.4472818]

[52] Barker J, Hannay P, Szewczyk P. Using traffic analysis to identify the second generation onion router. In: Proc. of the 9th IFIP Int'l Conf. on Embedded and Ubiquitous Computing (EUC). IEEE, 2011. 72–78. [doi: 10.1109/EUC.2011.76]

[53] Dixon L, Ristenpart T, Shrimpton T. Network traffic obfuscation and automated Internet censorship. IEEE Security & Privacy, 2016,14(6):43–53. [doi: 10.1109/MSP.2016.121]

[54] Song M, Xiong G, Li Z, *et al*. A de-anonymize attack method based on traffic analysis. In: Proc. of the Int'l ICST Conf. on Communications and NETWORKING in China. IEEE, 2014. 455–460. [doi: 10.1109/ChinaCom.2013.6694639]

[55] Almubayed A, Hadi A, Atoum J. A model for detecting Tor encrypted traffic using supervised. Machine Learning, 2015,7(7): 10–23.

[56] Shahbar K, Zincir-Heywood AN. Benchmarking two techniques for Tor classification: Flow level and circuit level classification. In: Proc. of the Computational Intelligence in Cyber Security. IEEE, 2015. 1–8. [doi: 10.1109/CICYBS.2014.7013368]

[57] Lashkari AH, Gil GD, Mamun MSI, *et al*. Characterization of Tor traffic using time based features. In: Proc. of the Int'l Conf. on Information Systems Security and Privacy. 2017. 253–262.

[58] Deng Z, Qian G, Chen Z, *et al*. Identifying Tor anonymous traffic based on gravitational clustering analysis. In: Proc. of the Int'l Conf. on Intelligent Human-Machine Systems and Cybernetics. IEEE, 2017. [doi: 10.1109/IHMSC.2017.133]

[59] Hodo E, Bellekens X, Iorkyase E, *et al*. Machine learning approach for detection of nonTor traffic. Journal of Cyber Security and Mobility, 2017,6(2):171–194.

[60] Lotfollahi M, Shirali R, Siavoshani MJ, *et al*. Deep packet: A novel approach for encrypted traffic classification using deep learning. 2017. http://xueshu.baidu.com/s?wd=paperuri%3A%282bdf662742490a13dadbc5c37fbd0aff%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Farxiv.org%2Fpdf%2F1709.02656&ie=utf-8&sc_us=7953292138050080248

[61] Tor-nonTor dataset. http://www.unb.ca/cic/datasets/tor.html

[62] Zhao GF, Chao-Ming JI, Chuan XU. Survey of techniques for Internet traffic identification. Journal of Chinese Computer Systems, 2010,31(8):1514–1520 (in Chinese with English abstract).

[63] Wang J, He H, Luo X, *et al*. Network traffic classification based on ensemble learning and co-training. Science in China, 2009, 52(2):338–346.

[64] Lü B, Liao Y, Xie HY. Survey on attack technologies to Tor anonymous network. Journal of CAEIT, 2017,12(1):14–19 (in Chinese with English abstract).

[65] Berthold O, Federrath H, Köhntopp M. Project anonymity and unobservability in the Internet. In: Proc. of the 10th Conf. on Computers, Freedom and Privacy: Challenging the Assumptions. ACM, 2000. 57–65.

[66] Agrawal D, Kesdogan D. Measuring anonymity: The disclosure attack. IEEE Security & Privacy, 2003,99(6):27–34. [doi: 10.1109/MSECP.2003.1253565]

[67] Danezis G. Statistical disclosure attacks. In: Security and Privacy in the Age of Uncertainty. Boston: Springer-Verlag, 2003. 421–426.

[68] Qin Y, Huang D, Li B. STARS: A statistical traffic pattern discovery system for MANETs. IEEE Trans. on Dependable and Secure Computing, 2014,11(2):181–192. [doi: 10.1109/TDSC.2013.33]

[69] Mallesh N, Wright M. An analysis of the statistical disclosure attack and receiver-bound cover. Computers & Security, 2011,30(8): 597–612. [doi: 10.1016/j.cose.2011.08.011]

[70] Bagai R, Lu H, Tang B. On the sender cover traffic countermeasure against an improved statistical disclosure attack. In: Proc. of the IEEE/IFIP Int'l Conf. on Embedded and Ubiquitous Computing. IEEE, 2011. 555–560. [doi: 10.1109/EUC.2010.90]

[71] Herrmann D, Wendolsky R, Federrath H. Website fingerprinting: Attacking popular privacy enhancing technologies with the multinomial Naïve-Bayes classifier. In: Proc. of the CCS 2009, Cloud Computing Security Workshop. 2009. 31–42. [doi: 10.1145/1655008.1655013]

[72] Murdoch SJ, Danezis G. Low-Cost traffic analysis of Tor. In: Proc. of the IEEE Symp. on Security & Privacy. IEEE, 2005. 183–195. [doi: 10.1109/SP.2005.12]

[73] Fusenig V, Staab E, Sorger U, *et al*. Slotted packet counting attacks on anonymity protocols. In: Proc. of the Australasian Conf. on Information Security. Australian Computer Society, Inc., 2009. 53–60.

[74] Murdoch SJ. Hot or not: Revealing hidden services by their clock skew. In: Proc. of the 13th ACM Conf. on Computer and Communications Security. ACM, 2006. 27–36.

[75] Weinberg Z, Wang J, Yegneswaran V, *et al*. StegoTorus: A camouflage proxy for the Tor anonymity system. In: Proc. of the 2012 ACM Conf. on Computer and Communications Security. ACM, 2012. 109–120.

[76] Biryukov A, Pustogarov I, Weinmann RP. Trawling for tor hidden services: Detection, measurement, deanonymization. In: Proc. of the 2013 IEEE Symp. on Security and Privacy (SP). IEEE, 2013. 80–94 .[doi: 10.1109/SP.2013.15]

[77] Liberatore M, Levine BN. Inferring the source of encrypted HTTP connections. In: Proc. of the ACM Conf. on Computer and Communications Security. ACM, 2006. 255–263.

[78] Wang T, Cai X, Nithyanand R, *et al*. Effective attacks and provable defenses for Website fingerprinting. In: Proc. of the USENIX Security Symp. 2014. 143–157.

[79] Kwon A, AlSabah M, Lazar D, *et al*. Circuit fingerprinting attacks: Passive deanonymization of tor hidden services. In: Proc. of the USENIX Security. 2015. 20.

[80] Hayes J, Danezis G. *k*-Fingerprinting: A robust scalable Website fingerprinting technique. In: Proc. of the USENIX Security Symp. 2016. 1187–1203.

[81] Zhuo Z, Zhang Y, Zhang Z, *et al*. Website fingerprinting attack on anonymity networks based on profile hidden Markov model. IEEE Trans. on Information Forensics and Security, 2017. [doi: 10.1109/TIFS.2017.2762825]

[82] Juarez M, Afroz S, Acar G, *et al*. A critical evaluation of Website fingerprinting attacks. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2014. 263–274. [doi: 10.1145/2660267.2660368]

[83] Guo XJ, Cheng G, Zhu CG, *et al*. Progress in research on active network flow watermark. Journal on Communications, 2014,35(7): 178–192 (in Chinese with English abstract). [doi: 1000-436X(2014)07-0178-15]

[84] Pyun YJ, Park YH, Wang X, *et al*. Tracing traffic through intermediate hosts that repacketize flows. In: Proc. of the INFOCOM the 26th IEEE Int'l Conf. on Computer Communications. IEEE, 2007. 634–642. [doi: 10.1109/INFCOM.2007.80]

[85] Houmansadr A, Borisov N. SWIRL: A scalable watermark to detect correlated network flows. In: Proc.of the Network and Distributed System Security Symp., NDSS 2011. San Diego: DBLP, 2011.

[86] Kiyavash N, Houmansadr A, Borisov N. Multi-Flow attacks against network flow watermarking schemes. In: Proc. of the Conf. on Security Symp. USENIX Association, 2008. 307–320.

[87] Luo J, Wang X, Yang M. An interval centroid based spread spectrum watermarking scheme for multi-flow traceback. Journal of Network and Computer Applications, 2012,35(1):60–71. [doi: 10.1016/j.jnca.2011.03.003]

[88] Wang X, Luo J, Yang M. A double interval centroid-based watermark for network flow traceback. In: Proc. of the Int'l Conf. on Computer Supported Cooperative Work in Design. IEEE, 2010. 146–151. [doi: 10.1109/CSCWD.2010.5471985]

[89] Wang X, Luo J, Yang M, *et al*. A novel flow multiplication attack against Tor. In: Proc. of the Int'l Conf. on Computer Supported Cooperative Work in Design. IEEE Computer Society, 2009. 686–691. [doi: 10.1109/CSCWD.2009.4968138]

[90] Abbott TG, Lai KJ, Lieberman MR, *et al*. Browser-Based attacks on Tor. In: Proc. of the Int'l Symp. on Privacy Enhancing Technologies, PET 2007. Ottawa: DBLP, 2007. 184–199.

[91] Dainotti A, Pescape A, Ventre G. A packet-level traffic model of starcraft. In: Proc. of the Int'l Workshop on Hot Topics in Peer-to-Peer Systems, Hot-P2P. IEEE, 2005. 33–42. [doi: 10.1109/PTPSYS.2005.4]

[92] Murdoch SJ, Danezis G. Low-Cost traffic analysis of Tor. In: Proc. of the 2005 IEEE Symp. on Security and Privacy. IEEE, 2005. 183–195. [doi: 10.1109/SP.2005.12]

[93] Mclachlan J, Hopper N. On the risks of serving whenever you surf: Vulnerabilities in Tor's blocking resistance design. In: Proc. of the ACM Workshop on Privacy in the Electronic Society. ACM, 2009. 31–40. [doi: 10.1145/1655188.1655193]

[94] Winter P, Lindskog S. How China is blocking Tor. In: Proc. of the USENIX Workshop on Free and Open Communications on the Internet (FOCI). 2012.

[95] Ensafi R, Winter P, Mueen A, *et al*. Analyzing the Great Firewall of China over space and time. Proc. on Privacy Enhancing Technologies, 2015,2015(1):61–76. [doi: 10.1515/popets-2015-0005]

[96] Tan J, Chen XS, Min DU, *et al*. Internet traffic identification algorithm based on adaptive BP neural network. In: Proc. of the Workshop on Intelligent Information Technology Applications. IEEE, 2012. 151–154. [doi: 10.3969/j.issn.1001-0548.2012.04.020]

[97] Ling Z, Luo J, Wu K, *et al*. TorWard: Discovery, blocking, and traceback of malicious traffic over Tor. IEEE Trans. on Information Forensics & Security, 2015,10(12):2515–2530. [doi: 10.1109/TIFS.2015.2465934]

附中文参考文献:

[14] 黄艳梅,林艳华.跨国网络犯罪愈演愈烈中国东盟谋求携手打击.2015. http://www.12377.cn/txt/2015-09/15/content_8235494.htm

[17] 工业和信息化部.工业和信息化部关于清理规范互联网网络接入服务市场的通知.http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757020/c5471946/content.html

[18] 何高峰,杨明,罗军舟,张璐.Tor 匿名通信流量在线识别方法.软件学报,2013(3):540–556. http://www.jos.org.cn/1000-9825/4253.htm [doi: 10.3724/SP.J.1001.2013.04253]

[32] 李响.基于 Meek 的 Tor 匿名通信识别方法的研究和实现[硕士学位论文].北京:北京交通大学,2016.

[33] 谭庆丰,时金桥,方滨兴,等.匿名通信系统不可观测性度量方法.计算机研究与发展,2015,52(10):2373–2381. [doi: 10.3969/j.issn.1001-0548.2012.04.020]

[39] 何永忠,陈美玲.基于协议的拟态研究综述.北京交通大学学报,2016,40(5):1–8. [doi:10.11860/j.issn.1673-0291.2016.05.001]

[42] 吴倩.基于 DPI 与 DFI 的流量识别与控制系统的设计与实现[博士学位论文].成都:电子科技大学,2013.

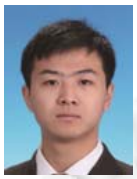
[44] 鲁刚,张宏莉,叶麟.P2P 流量识别.软件学报,2011,22(6):1281−1298. http://www.jos.org.cn/1000-9825/3995.htm [doi: 10.3724/SP.J.1001.2011.03995]

[48] 吴震,刘兴彬,童晓民.基于信息熵的流量识别方法.计算机工程,2009,35(20):115−116.

[62] 赵国锋,吉朝明,徐川.Internet 流量识别技术研究.小型微型计算机系统,2010,31(8):1514–1520. [doi:1000-1220(2010)08-1514-07]

[64] 吕博,廖勇,谢海永.Tor 匿名网络攻击技术综述.中国电子科学研究院学报,2017,12(1):14–19. [doi:10.3969/j.issn.1673-5692.2017.01.003]

[83] 郭晓军,程光,朱琛刚,等.主动网络流水印技术研究进展.通信学报,2014,35(7):178–192. [doi:10.3969/j.issn.1000-436x.2014.07.022]



姚忠将(1988－),男,山东聊城人,博士生,主要研究领域为流量识别与追踪,区块链,隐私保护,机器学习.

葛敬国(1973－),男,博士,研究员,博士生导师,主要研究领域为软件定义网络,网络虚拟化,云计算.

张潇丹(1983－),女,博士,副研究员,主要研究领域为未来网络实验环境,网络虚拟化及软件定义网络,新型网络技术测量分析与评估.

郑宏波(1977－),男,工程师,主要研究领域为软件定义网络,网络虚拟化,云计算.

邹壮(1993－),男,硕士生,主要研究领域为软件定义网络,网络虚拟化,云计算.

孙焜焜(1995－),男,硕士生,主要研究领域为软件定义网络.

许子豪(1995－),男,硕士生,主要研究领域为软件定义网络,网络功能虚拟化.