软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(3):540−556 [doi: 10.3724/SP.J.1001.2013.04253]

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

Tor 匿名通信流量在线识别方法*

何高峰, 杨 明, 罗军舟, 张 璐

(东南大学 计算机科学与工程学院,江苏 南京 211189)

通讯作者: 何高峰, E-mail: hegaofeng@seu.edu.cn

摘 要: 匿名通信技术的滥用给网络监管带来了新的挑战.有效识别出匿名通信流量,是阻止该类技术滥用的前提,具有重要的研究意义和应用价值.现有研究工作侧重于匿名通信关系的确认,无法用于匿名通信流量的识别和阻塞.针对这个问题,围绕广泛使用的 Tor 匿名通信系统,深入分析运行机制,归纳总结其流量特征.在此基础上,分别提出基于 TLS 指纹和基于报文长度分布的 Tor 匿名通信流量识别方法.对两种识别方法的优缺点和适用性进行了详细分析和讨论,并通过 CAIDA 数据集和在线部署对识别方法进行了验证.实验结果表明,基于 TLS 指纹和基于报文长度分布的识别方法均能有效识别出 Tor 匿名通信流量.

关键词: 匿名通信;Tor;流量识别;TLS 指纹;报文长度分布

中图法分类号: TP393 **文献标识码**: A

中文引用格式: 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,2013,24(3):540−556. http://www.jos.org.cn/1000-9825/4253.htm

英文引用格式: He GF, Yang M, Luo JZ, Zhang L. Online identification of Tor anonymous communication traffic. Ruanjian Xuebao/Journal of Software, 2013,24(3):540−556 (in Chinese). http://www.jos.org.cn/1000- 9825/4253.htm

Online Identification of Tor Anonymous Communication Traffic

HE Gao-Feng, YANG Ming, LUO Jun-Zhou, ZHANG Lu

(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Corresponding author: HE Gao-Feng, E-mail: hegaofeng@seu.edu.cn

Abstract: Abuse of anonymous communication systems has introduced new challenges into network administration. The effective identification of anonymous communication traffic is a prerequisite to prevent such abuse; thus, this is fundamentally important for both theoretical researches and practical applications. Existing researches mainly focus on the confirmation of anonymous communication relationship and cannot be used to identify and block anonymous communication traffic. To solve this problem, the operation mechanism is deeply analyzed and traffic characteristics are summarized for the widely used Tor anonymous communication system. On this basis, a TLS fingerprint-based and packet-size distributions based methods are proposed to identify Tor anonymous communication traffic, respectively. The advantages, disadvantages and applicability of these two methods are analyzed and discussed in detail, and are validated by CAIDA dataset and online deployment. Experimental results prove that both methods are effective in identifying Tor anonymous communication traffic.

Key words: anonymous communication; Tor; traffic identification; TLS fingerprint; packet size distribution

现有的匿名通信系统是建立在 Internet 之上、利用 Mix 技术隐藏网络通信中发送方与接收方的身份信息

* 基金项目: 国家重点基础研究发展计划(973)(2010CB328104); 国家自然科学基金(61272054, 61202449, 61070161, 61003257, 60903162); 国家科技支撑计划(2010BAI88B03, 2011BAK21B02); 国家核高基科技重大专项(软件类)项目(2010ZX01044-001-001); 高等学校博士点专项科研基金(20110092130002); 江苏省自然科学基金(BK2008030); 江苏省网络与信息安全重点实验室资助项目(BM2003201); 教育部计算机网络与信息集成重点实验室(东南大学)资助项目(93K-9)

收稿时间: 2011-07-12; 定稿时间: 2012-04-27

(如 IP 地址)以及双方通信关系的一种覆盖网络,由众多中继节点构成,用户发送的报文经数据加密、中继节点混淆(报文延迟、乱序、报文填充等)和多次转发后到达最终目的节点.匿名通信系统正是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.根据传输延时可划分为高延时匿名通信系统[1,2]和低延时匿名通信系统[3,4].高延时匿名通信系统仅适用于对实时性要求较低的网络应用,如电子邮件等,而低延时系统则支持诸如网页浏览、文件下载、即时通信(如 MSN,ICQ)等延时敏感类网络应用.由于现有的网络应用大多是低延时业务,而且低延时匿名通信系统同样可以实现高延时匿名通信系统的功能,因而本文的研究主要针对低延时匿名通信系统进行.为行文简洁,在不引起混淆时,下文简称低延时匿名通信系统为匿名通信系统.

匿名通信系统在保护网络用户隐私信息的同时也可能为不法分子所利用,为其实施网络犯罪增加掩护,阻碍审查机构对其的监控与追踪.因此,目前网络安全领域的一个研究热点便是匿名通信追踪技术,即通过还原匿名用户间的通信关系为匿名网络犯罪的追踪提供相应的技术手段.然而,仅能确定通信双方间的通信关系并不能满足某些特定的网络监管需求.如在机密部门内部网络环境中,网络管理员并不需要知道内部员工通过匿名通信系统访问了哪些限制网站,而应当及时识别并阻断匿名通信流量,从根本上消除借助匿名通信系统泄漏机密的可能.与此同时,在识别出匿名通信流量的基础上,还可执行其他网络监管策略,如决定是否执行匿名通信追踪、匿名通信内容分析等.匿名通信流量的识别是后续操作的基础,有着重要的研究意义和应用价值,针对匿名通信流量的识别方法研究已成为一项紧迫的网络监管任务.

针对上述需求,本文围绕目前使用最为广泛的Tor匿名通信系统[3]深入分析其运行机制,包括TLS连接和转发链路的建立、数据处理和信元(cell)封装流程,得出其 TLS 连接和报文长度分布特征.在此基础上,分别提出基于 TLS 指纹和基于报文长度分布的 Tor 匿名通信流量识别方法.基于 TLS 指纹的识别方法将{密码套件,数字证书}作为抽取 Tor 流量 TLS 指纹特征的依据,首先判断待识别流是否为 TLS 流量,其次进一步判断其密码套件和数字证书等特征是否匹配 Tor 的 TLS 指纹,从而判断出该流为 Tor 流量.基于报文长度的识别方法以报文长度分布为识别特征,采用支持向量机分类算法识别 Tor 流量.首先统计出 Tor 流量的报文长度分布特征,为降低特征维数、加快识别速度,选取一组特定的报文长度作为最终统计对象,形成 Tor 流量学习样本;然后统计非 Tor 流量中相同长度报文的分布特征,将其作为支持向量机的其他类型流量的学习样本.学习阶段根据学习样本生成判别函数,检测阶段则将待识别流的报文长度分布特征带入判别函数,若判别结果为 1,则判断该流为 Tor 流量;否则,判断为其他类型流量.

本文第 1 节简要介绍匿名通信系统 Tor 的基本架构以及匿名通信追踪和网络流量识别方面的相关研究工作.第 2 节深入分析 Tor 的运行机制,详细描述其匿名转发链路的建立、数据处理和信元封装流程以及报文长度分布特征.在此基础上,第 3 节提出 Tor 流量识别方法,包括基于 TLS 指纹和基于报文长度分布的识别方法,以及对这两种方法的分析与讨论.第 4 节为实验部分,包括数据分析、离线测试和在线识别.最后一节总结全文.

1 相关工作

Chaum 于 1981 年首先提出 Mix 技术和匿名通信的概念[5],后续匿名通信技术的发展均以此作为基础.目前,实用的匿名通信系统主要有 Tor[3]和 JAP[4]等.Tor 能支持所有基于 TCP 协议的上层应用;而 JAP 目前仅支持 HTTP 应用,并且 JAP 客户端程序在下载中继节点信息和与中继节点交互的起始阶段采用明文传输[6],易于识别,因而本文以 Tor 为分析和识别对象.本节首先对 Tor 匿名通信系统的系统架构作简要介绍,然后说明匿名通信追踪技术和网络流量识别技术的相关研究工作.

Tor 作为第 2 代洋葱路由匿名通信系统,是目前最为流行、应用最为广泛的匿名通信系统.当前,Tor 系统中约有 2 400 个 OR(Tor 中继节点被称为 onion router)节点(https://metrics.torproject.org/network.html),同时在线用户数达到 690 000(https://metrics.torproject.org/users.html).Tor 适用于所有基于 TCP 协议的应用,其整体系统架构如图 1 所示.当 Alice 通过 Tor 访问相关网络资源,如 Web 站点时,首先从目录服务器处下载所有的 OR 节点信息,然后根据各节点所公告的带宽、在线时间长短和设定的出口访问策略等因素选择 3 个节点分别作为入口节

点(entry node)、中间节点(middle node)和出口节点(exit node).随后,Alice 与 Web 间通信的所有报文都将通过该 3 跳路径转发.

匿名通信系统实现了网络通信关系的隐藏,现阶段对匿名通信关系的确认,即匿名通信追踪技术研究是一个重要方向.文献[7]针对低延时匿名系统在转发数据时无法完全消除流中数据包时间特征的缺陷,提出了基于包抵达时间的流水印技术;文献[8]使用直序扩频(direct sequence spread spectrum)技术将水印信号扩频后通过调制流速率的方式嵌入通信流中,有效地提高了流水印的检测率;文献[9]通过调制 Tor 节点一次发送信元数量的方法,在 Tor 的协议层嵌入水印,以确认发送者和接收者间的通信关系;文献[10]以时间间隔为水印载体,提出一种改进的双时隙质心匿名通信追踪技术;文献[11]以流时隙质心为水印载体,通过延迟报文方式嵌入水印信息,首次提出一种流相关的匿名通信追踪技术.

与匿名通信追踪技术相比,匿名通信流量识别是根据一定流特征识别出网络中匿名通信流量的技术.当前,Tor 目录服务器和各 OR 节点的 IP 地址均公开,可直接根据报文 IP 地址识别出 Tor 流量,从而阻止 Tor 的使用.但 Tor 已采取相应的反阻塞措施,Tor 用户可配置自身成为一种 IP 地址并不公开的 OR 节点,即 Bridge 节点.其他用户可直接与 Bridge 节点通信,从 Bridge 节点处下载所有其他公开的 OR 节点信息,并以 Bridge 节点作为自己的入口节点,如图 1 中虚线所示.由于 Bridge 节点的 IP 地址并不公开,只能通过邮件或网页方式获得,而且 Bridge 节点的发布有着各种限制[12],网络管理者几乎不可能获取所有的 Bridge 节点信息,因此无法仅借助 IP 地址来识别出所有 Tor 流量.

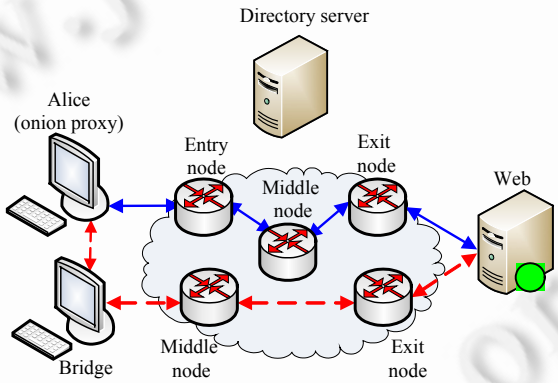


Fig.1 Tor system architecture

图 1 Tor 系统架构

目前,对匿名通信流量识别的研究尚未引起足够重视,相关研究工作较少.与匿名通信流量识别类似的研究工作有 P2P 流量识别和网络异常流量检测等.P2P 流量识别采用端口识别[13]、应用层签名[14]和统计特征[15]等技术手段识别出网络中的 P2P 流量,但匿名通信协议的设计和实现中均未使用固定端口,且所有报文都已加密,因而端口识别和应用层签名技术并不适用于识别匿名通信流量.特征统计仅针对 P2P 协议和其系统实现,无法直接用于匿名通信流量识别,但其研究思路有重要参考价值.网络异常流量检测并不针对某类具体应用,而是从安全角度出发检测网络中可能的异常流量,如蠕虫爆发、DDoS 攻击和网络设备故障等.文献[16]指出,网络链路中正常的流分布符合短时间均衡模型(ASTUTE),当有异常流量出现时,此模型便不再成立,因而检测出异常流量;文献[17]利用异常的通信流、目的地址和目的端口数量等通信特征来检测网络中的蠕虫病毒.网络异常流量检测通常以网络整体为考察对象,其特征提取和检测算法并不适用于检测匿名通信流量.本文在对 Tor 匿名通信系统深入分析的基础上,提出基于 TLS 指纹和基于报文长度分布的 Tor 流量识别方法.本文还对两种识别方法的优缺点和适用性进行了分析讨论.

2 Tor 运行机制分析

本节对 Tor 匿名通信系统的运行机制进行深入分析,包括匿名转发链路的建立、数据处理和信元封装流程以及报文长度分布特征,归纳总结出 Tor 流量的特点,为对其的识别提供可靠依据.

2.1 转发链路的建立

如图 1 所示,Alice 处的 OP(onion proxy)程序启动后,根据 OR 节点的带宽大小、在线时间长短和设定的出口策略等因素选择 3 个 OR 节点,分别作为入口节点、中间节点和出口节点构建匿名转发链路.转发链路之间均采用 TLS 加密,其建立流程如图 2 所示.OP 首先与入口节点建立 TLS 连接,然后发送 Create 请求,与入口节点协商会话对称密钥,该对称密钥由 DH(diffie-Hellman)密钥交换协议生成,并使用 RSA 加密算法对 DH 交换参数进行加密处理.协商成功后,OP 通过入口节点,继续与中间节点和出口节点间进行类似交互,建立 3 跳的匿名转发链路.当 OP 传输数据时,数据经入口节点、中间节点和出口节点层层加/解密和转发操作,最终到达目的端.

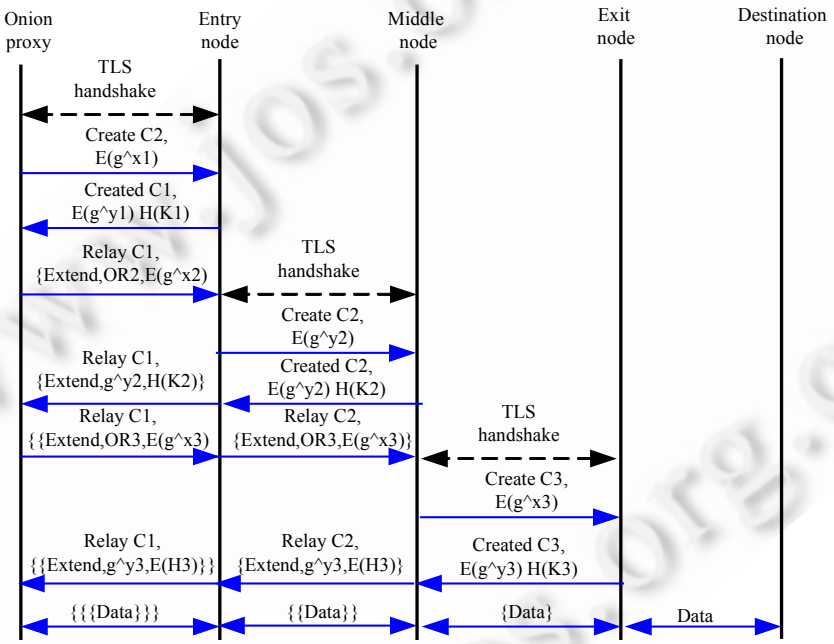


Fig.2 Procedure of circuit creation on Tor

图 2 Tor 匿名转发链路建立流程

由上述分析可知,Tor 采用 TLS 安全协议对转发链路进行加密,其 TLS 连接建立过程如图 3 所示.首先,客户端发送 Client Hello 报文,此报文包括有客户端支持的密码套件和用于产生会话密钥的随机数.服务器端接收到 Client Hello 报文后,发送 Server Hello,Certificate,Server Key Change 和 Server Hello Done 这 4 个报文.其中,Server Hello 报文包含有服务器端产生的随机数和选择的密码套件,协商使用 DH 对称密钥交换协议生成预主密钥,签名算法选用 RSA,会话阶段的对称加密算法为 AES;Certificate 报文为服务器的证书信息,该证书由 Tor 程序每隔 2 小时更新 1 次;Server Key Change 报文包含 DH 密钥交互协议中服务器端的公开参数部分和对应的签名;Server Hello Done 表示服务器端发送结束.客户端在接收到 Server Hello Done 报文后,生成预主密钥和会话密钥,DH 密钥交互协议中客户端的公开参数由 Client Key Change 报文发送,同时发送 Change Cipher Spec 报文,表明后续报文均将用刚刚协商的会话密码进行加密.Finished 报文则包含了客户端整个连接过程的校验,用于握手过程验证.服务器端根据预主密钥、Client Hello 和 Server Hello 报文中的随机数生成相同的会话密钥,发送 Change Cipher Spec 和 Finished 报文,至此,TLS 连接建立结束.

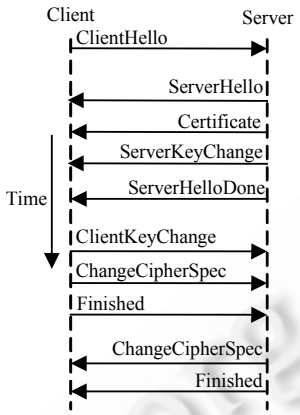


Fig.3 Procedure of TLS handshake

图 3 TLS 连接建立流程

在此过程中,OR 节点发送的 Server Hello 报文中选择的密码套件固定为 DHE_RSA_WITH_AES_256_SHA,DHE_RSA_WITH_AES_128_SHA,EDH_RSA_DES_192_CBC3_SHA 这 3 种选择之一.Certificate 报文中的 X.509 证书由 OR 节点每隔 2 小时重新生成,该证书包括的主要证书项和对应的赋值可见表 1.其中,颁发机构和拥有者名称中的 *xyz* 表示长度为 8 位~20 位的数字和英文字母的随机组合.而目前可用的 TLS 密码套件约有 50 种(http://www.openssl.org/docs/apps/ciphers.html#TLS_v1_0_cipher_suites_),证书有效时间一般为 1 年~5 年(http://www.thawte.com/ssl/web-server-ssl-certificates/index.html).颁发机构和拥有者名称具有一定含义,因而 Tor 在建立 TLS 连接时选择的密码套件、X.509 数字证书中的序列号、颁发机构和拥有者名称、起效时间和失效时间等特征明显,可用于识别 Tor 流量.

Table 1 Tor certificate structure

表 1 Tor 证书格式

证书项	赋值
版本号	2
序列号	证书起效时间
签名算法	SHAwithRSA
公钥信息	证书公钥
数字签名	证书签名
颁发机构名称	www.*xyz*.net
拥有者名称	www.*x′y′z′*.net
证书起效时间	生成证书时的本地当前时间
证书失效时间	起效时间+2 小时

2.2 数据处理和信元封装流程

转发链路建立成功后,OP 程序监听本地端口(默认为 9050).其他应用程序通过配置 Socks 代理(127.0.0.1:9050)连接至 OP.当 OP 接收到应用程序的数据后,一次整取 498 字节,不足 498 字节长度的则进行填充,并用转发链路上 3 跳节点的会话密钥对该数据进行层层加密;最后,添加头部字段构成 512 字节的 Tor 信元结构,交由 TLS 层再次加密并传输.

Tor 的数据处理和转发功能均由 libevent(http://monkey.org/~provos/libevent/)的读事件和写事件控制完成.具体的,OP 程序有一个输入缓冲区和一个输出缓冲区.应用程序通过 Socks 代理将数据写入、输入缓冲区中.当读事件被调度时,OP 从输入缓冲区中读取数据,一次最多读取 498 字节,然后利用出口节点、中间节点和入口节点的会话密钥对读取的数据层层加密,再增加信元头部字段,并将一个长度为 512 字节的完整信元写入信元

队列中.若输入缓冲区中还有数据,则继续上述操作,直至读取完所有数据,并注册写事件.当写事件被调度时,首先将输出缓冲区中的信元尽可能多地写入 TLS 写缓冲区中,然后将信元队列中的信元尽可能多地写入输出缓冲区中.即一个信元发送至网络需要两次写事件,一次写事件将信元写入输出缓冲区中,另一次写事件将信元从输出缓冲区中写入 TLS 层加密并发送.Tor 信元结构如图 4 所示.图 4(a)为 Tor 命令控制信元结构,图 4(b)为数据转发信元结构.

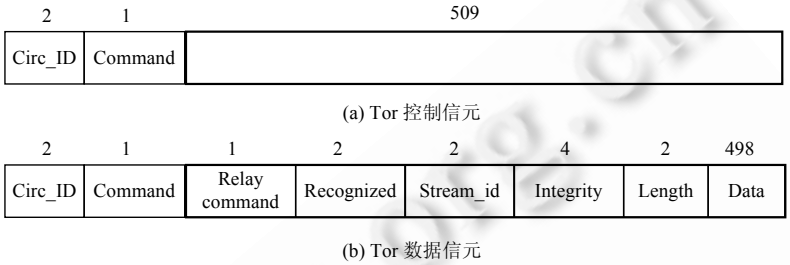


Fig.4 Tor cell format

图 4 Tor 信元结构

OP 的报文处理流程如图 5 所示.

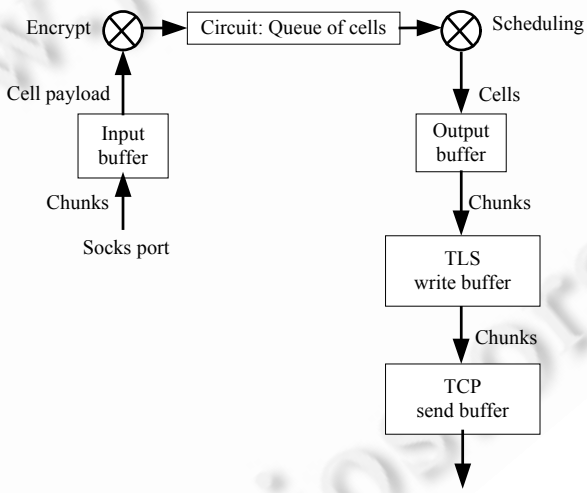


Fig.5 Procedure of cell processing on OP

图 5 OP 节点报文处理流程

转发链路上的 OR 节点对数据的处理流程与 OP 略有不同,OR 节点从 TLS 读缓冲区中读取数据至输入缓冲区中,并且每次从输入缓冲区中读取一个信元长度的数据,即每次读取 512 字节.根据路径方向执行完加密或解密操作后根据信元的 Circ_ID 字段将该信元放入相应队列中,OR 节点有多个信元队列.具体流程如图 6 所示.

在图 5 和图 6 中,输出缓冲区中的信元写入 TLS 写缓冲区中,加密后经 TCP 协议传输.Tor 借由 OpenSSL (http://www.openssl.org/)实现 TLS 层加密.TLS 报文由 TLS 头部、加密后的数据和消息验证码 MAC 以及填充字段组成,具体格式如图 7(a)所示.但 Tor 会对每个 TLS 报文增加一个空的 TLS 记录,因而最终的 TLS 报文结构如图 7(b)所示.

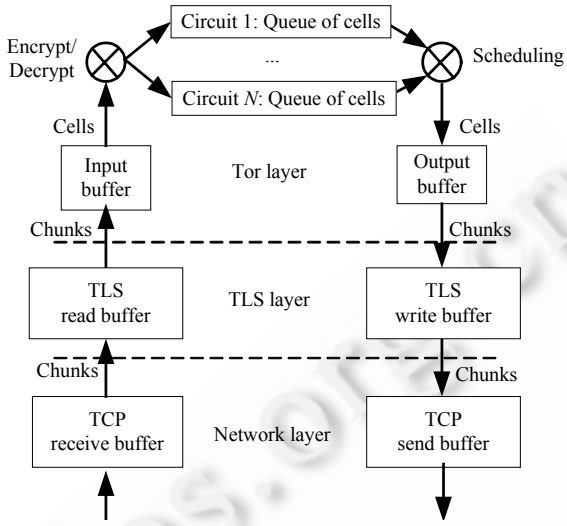


Fig.6 Procedure of cell processing on OR

图 6 OR 节点报文处理流程

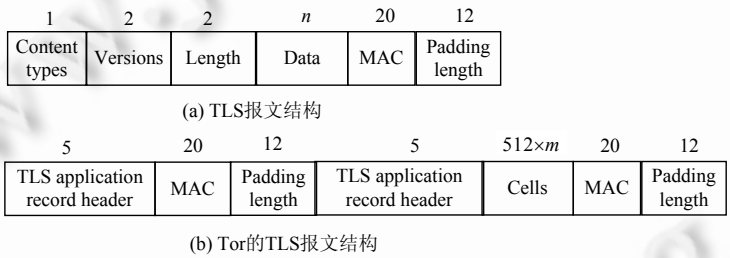


Fig.7 Tor TLS packet format

图 7 Tor TLS 报文结构

2.3 报文长度分布特征

由第 2.2 节的描述可知,TCP 发送缓冲区中包含一个或多个 TLS 报文,每个 TLS 报文中有一个或多个 Tor 信元报文.这些 TLS 报文经一个或多个 TCP 报文发送.设 TCP 发送缓冲区中共有 k 个 TLS 报文,而 k 个 TLS 报文中共有 m 个 Tor 信元报文,则 TCP 发送缓冲区中数据总长度为

$$512\times m+(5+20+12)\times 2\times k=512\times m+72\times k \tag{1}$$

记网络最大报文段长度(maximum segment size)为 M,则网络层观察到的典型报文长度(若 TCP 报文通告的接收窗口小于 M 值,则可能会形成具有其他长度值的报文,但此情形并不常见)为

$$\{(512\times m+72\times k) \bmod M\} \tag{2}$$

由于其他应用同样会产生长度正好为 M 的报文,因而不将 M 值统计为 Tor 流量的典型报文长度.其中,m 和 k 的取值范围由 TLS 写缓冲和 TCP 发送缓冲区长度所决定.设 TLS 写缓冲长度为 W,TCP 发送缓冲长度为 S,则有:

$$1\leqslant m\leqslant\left\lfloor\frac{W}{512}\right\rfloor \tag{3}$$

$$1\leqslant k\leqslant\left\lfloor\frac{S}{586}\right\rfloor \tag{4}$$

并应同时满足:

$$512\times m+72\times k\leqslant S \tag{5}$$

公式(4)中的 586 对应于 k=1,m=1 时 TCP 发送缓冲区中的数据长度,即只有一个 TLS 报文,且该 TLS 报文中只包含一个信元报文.

由公式(2)可求出 Tor 流量中典型报文长度,如当 M=1360B,W=16KB,S=16KB 时,公式(2)共有 363 种不同计算结果值,典型报文长度为:

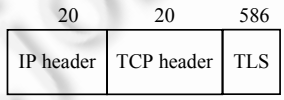
(i) m=1,k=1:报文长度为 586 字节,对应 TCP 发送缓冲区中有一个 TLS 报文,且该 TLS 报文中包含一个信元报文(如图 8(a)所示);

(ii) m=2,k=1:报文长度为 1 098 字节,对应 TCP 发送缓冲区中有一个 TLS 报文,且该 TLS 报文中包含两个信元报文(如图 8(b)所示);

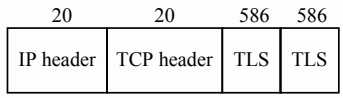
(iii) m=2,k=2:报文长度为 1 172 字节,对应 TCP 发送缓冲区中有两个 TLS 报文,每个 TLS 报文中包含一个信元报文(如图 8(c)所示);

(iv) m=3,k=3:报文长度为 410 字节,对应 TCP 发送缓冲区中有 3 个 TLS 报文,每个 TLS 报文中包含一个信元报文(如图 8(d)所示);

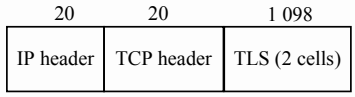
(v) m 和 k 为其他值时,报文长度可为 110,262,774 等字节大小.



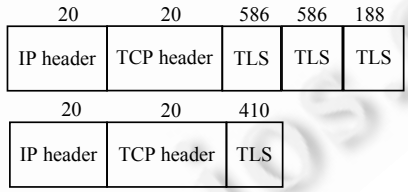
(a) IP报文中有1个Tor TLS报文: 586字节



(b) IP报文中有2个Tor TLS报文: 1 172字节



(c) IP报文中有1个Tor TLS报文: 1 098字节



(d) IP报文中有3个Tor TLS报文: 1 360和410字节

Fig.8 Packet format of typical size

图 8 典型长度报文结构

在计算出 Tor 流量中典型报文长度的基础上,结合常见的、B/S 或 C/S 架构类型(关于非 B/S 和 C/S 架构类型的应用,如 P2P 的讨论见第 3.3 节)的 Tor 上层应用程序的流量特征,可以归纳出各典型长度报文出现的频率大小特点,进而得出 Tor 流量的报文长度分布特征.在 B/S 或 C/S 架构类型的应用中,通常客户端发出少量请求,而服务器端返回大量数据.客户端发出的请求报文数据量小,并且两次请求报文之间有一定时间间隔,因而其报文在通过 Tor 节点转发时不易发生排队现象,从而报文长度主要为公式(2)中对应的 m 和 k 取值较小时的计算结果值.其 Tor 流量的报文长度分布特征可归纳为少数典型长度报文大量出现.

由服务器端返回的大量数据在通过 Tor 节点转发时容易产生报文排队,且队列长度随数据大小的改变而变化,从而形成大量长度为 M 而尾部为多种其他长度的报文.由文献[18]可知,客户端接收到的 Tor 流量中报文长

度分布与其他应用相近,而由客户端发出的 Tor 流量报文长度分布特征具有特殊性,可作为其识别特征.

3 Tor 匿名通信流量识别

由第 2 节分析得知,可利用{密码套件,数字证书}和报文长度分布特征识别 Tor 流量.本节给出基于 TLS 指纹和基于报文长度分布识别方法的具体描述,并对两种不同方法进行分析比较和进一步讨论.

3.1 基于TLS指纹的Tor流量识别

由第 2.1 节得知,可从{密码套件,数字证书}中抽取 Tor 流量 TLS 指纹特征,如密码套件固定为 DHE_RSA_WITH_AES_256_SHA,DHE_RSA_WITH_AES_128_SHA,EDH_RSA_DES_192_CBC3_SHA 这 3 种选择之一;证书序列号等于证书起效时间;证书的颁发机构和拥有者名称结构固定,均为 www.xyz.net,且 xyz 的长度范围为 8~20;证书起效时间与当前时间接近,且有效时间长度为 2 小时.

基于 TLS 指纹的 Tor 流量识别方法可表示为如下步骤:

步骤 1:由 Client Hello 报文判断待识别流是否为 TLS 或 SSLV3 流量(为支持低版本的 OpenSSL,Tor 也会采用 SSL V3 协议进行链路层加密).若成立,执行步骤 2;否则,判断为其他类型流量;

步骤 2:判断 Server Hello 报文中密码套件是否为 DHE_RSA_WITH_AES_256_SHA,DHE_RSA_WITH_AES_128_SHA,EDH_RSA_DES_192_CBC3_SHA.若成立,执行步骤 3;否则,判断为其他类型流量;

步骤 3:提取 Certificate 报文中的证书序列号、颁发机构和拥有者名称、起效时间和失效时间;

步骤 4:判断颁发机构和拥有者名称是否满足 www.xyz.net 结构,且 xyz 的长度在 8~20 该范围内.若成立,执行步骤 5;否则,判断为其他类型流量;

步骤 5:将 ASCII 码形式的起效时间和失效时间转换为长整型数值,并判断证书序列号与起效时间是否相等.若成立,执行步骤 6;否则,判断为其他类型流量;

步骤 6:判断失效时间与起效时间的差值是否等于 2 小时,并且起效时间与当前时间差值的绝对值小于一定门限值,如 24 小时.若成立,判断为 Tor 流量;否则,判断为其他类型流量.

在步骤 6 中,考虑到时间不同步问题,故而判断起效时间与当前时间是否接近时设定一时间差门限值,且该门限值应大于 2 小时.

基于 TLS 指纹的识别方法仅需分析 TLS 流量中 Client Hello,Server Hello 和 Certificate 报文,识别速度快,能适用于在线识别.但若密码套件、数字证书等特征发生改变,识别方法需同步做出改变.基于报文长度分布的识别方法能有效克服上述缺陷,具体方法描述见第 3.2 节.

3.2 基于报文长度分布的Tor流量识别

Tor 流量报文长度分布特征可描述为少数典型长度报文大量出现,但由于上层应用的不同、Tor 读写事件调度的不确定性和网络环境的实时变化,具体到每条 Tor 流量时其报文长度分布可能各不相同,因而难以建立统一的报文长度分布线性模型.鉴于支持向量机[19]在解决小样本、非线性识别问题中的独特优势,本文采用支持向量机分类算法判断待识别流的报文长度分布是否满足 Tor 流量的分布特征.下文中称由客户端发出的流量为上行流量,接收到的流量为下行流量.

基于支持向量机的 Tor 流量识别方法由特征选取、学习和检测这 3 个阶段组成.在特征选取阶段,针对数据集中所有 Tor 上行流量,将典型长度报文按出现的频率由高到低排序,并选取频率和值大于设定的门限值,如 90%的前 l 位频率组成 Tor 流量的报文长度分布 $P=\{p_1,p_2,\ldots,p_l\}$.将长度等于 s_i 的报文称为特征报文,p_i 与报文长度 s_i 一一对应,p_i 计算为

$$p_i=\frac{n(s_i)}{n(A)} \tag{6}$$

其中,$n(A)$表示 Tor 上行流量中长度大于 0 的所有报文数量,$n(s_i)$表示报文长度为 s_i 的报文数量.

在确定特征报文后,统计每条 Tor 上行流量和其他类型流量中相同特征报文出现的频率,形成学习样本.学

习样本形式化地表示为

$$\langle 1,\{p_A^1,p_A^2,...,p_A^l\}\rangle \tag{7}$$

$$\langle -1,\{p_N^1,p_N^2,...,p_N^l\}\rangle \tag{8}$$

其中,1 和−1 表示样本类别,分别代表 Tor 流量和其他类型流量. p_A^i 表示 Tor 流量中长度为 s_i 的报文出现频率, p_N^i 表示其他类型流量中长度为 s_i 的报文出现频率.

对于公式(7)和公式(8)表示的样本数据,在学习阶段,支持向量机的学习目标是构造判别函数将样本数据尽可能正确分类.令 y_i=1 或 y_i=−1,P 表示流报文长度分布,$\phi(\cdot)$为非线性函数,共有 k 个学习样本.若存在分类超平面

$$w\cdot\phi(P)+b=0 \tag{9}$$

使得

$$\begin{aligned} &w\cdot\phi(P_i)+b\geqslant 1, y_i=1\\ &w\cdot\phi(P_i)+b\leqslant -1, y_i=-1, i=1,2,...,k \end{aligned} \tag{10}$$

公式(10)可统一表示为

$$y_i\cdot(w\cdot\phi(P_i)+b)\geqslant 1, i=1,2,\ldots,k \tag{11}$$

为解决线性不可分性问题,引入非负松弛变量ε_i,公式(11)进一步变换为

$$y_i\cdot(w\cdot\phi(P_i)+b)\geqslant 1-\varepsilon_i, i=1,2,\ldots,k \tag{12}$$

对于公式(12),支持向量机寻找最优超平面.

最优超平面问题描述为

$$\left.\begin{aligned} &\min_{w,b,\varepsilon}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{k}\varepsilon_i\\ &\text{s.t. } y_i(w\cdot\phi(P_i)+b)\geqslant 1-\varepsilon_i\\ &\varepsilon_i\geqslant 0, i=1,2,...,k \end{aligned}\right\} \tag{13}$$

其中,ε_i 表示松弛变量;C 表示边际系数,用来平衡最大间隔和最小分类误差.

采用拉格朗日乘子法将公式(13)变换为对偶问题:

$$\left.\begin{aligned} &\max_{\alpha}\left\{\sum_{i=1}^{k}\alpha_i-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}y_iy_j\phi(P_i)\cdot\phi(P_j)=\sum_{i=1}^{k}\alpha_i-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}y_iy_jH(P_i\cdot P_j)\right\}\\ &\text{s.t.}\quad 0\leqslant\alpha_i\leqslant C\\ &\sum_{i=1}^{k}\alpha_iy_i=0 \end{aligned}\right\} \tag{14}$$

其中,α_i 为拉格朗日乘子,$H(P_i\cdot P_j)=\phi(P_i)\cdot\phi(P_j)$为核函数.本文采用径向基函数为相应的核函数

$$H(P_i,P_j)=\exp(-\gamma\|P_i-P_j\|^2) \tag{15}$$

其中,γ为控制半径的正数.

由公式(14)求出支持向量集 SV 和对应的α_i,最终形成的判别函数为

$$y(P)=sign\left[\sum_{P_i\in SV}\alpha_iy_iH(P_i,P)+b\right] \tag{16}$$

其中,$sign(\cdot)$为符号函数,SV 为学习阶段形成的支持向量集,b 值可计算为

$$b=\frac{1}{|U|}\sum_{P_i\in U}\left[y_i-\sum_{P_j\in SV}\alpha_jy_jH(P_j,P_i)\right] \tag{17}$$

其中,U 为所有非边界支持向量集($0<\alpha_i<C$ 对应的支持向量集).

学习阶段为离线训练过程,形成判别函数后,在检测阶段,只需将待识别流 F 的特征报文分布 P 带入公式(16)中,若计算结果为 1,则判断 F 为 Tor 流量,否则为其他类型流量.

3.3 分析讨论

本节对基于TLS指纹和基于报文长度分布这两种Tor流量识别方法进行比较,进而分析两种识别方法的适用性特点.此外,还分析了P2P架构类型对Tor流量识别的影响.

基于TLS指纹的识别方法仅需要分析TLS握手阶段的Client Hello,Server Hello和Certificate报文,部署便捷,识别速度快.但其存在的缺陷是,如果Tor修改了其密码套件、证书时间或颁发机构名称等特征,该识别方法需同步做出修改,甚至无法识别出Tor流量.如:Tor可以将其证书起效时间和证书有效时间长度设定为任意值;随机生成颁发机构和拥有者名称等.因而在使用基于TLS指纹的识别方法时,需关注发布的最新Tor版本是否改变了对应的TLS指纹特征.

基于报文长度分布的识别方法需要有前期学习过程,增加了方法的实施复杂度;在线识别时需统计一定数量的报文,因而与基于TLS指纹的识别方法相比,其识别速度较慢.但该识别方法具有通用性.这是因为将应用层数据封装成等长度的信元是匿名通信系统设计的基本特征,如Tor的信元长度为512字节,JAP中的信元设计为998字节.采取此类设计是为了消除攻击者利用报文长度特征来确定通信关系的可能性.

如图9所示,如果匿名通信系统不对上层应用数据进行等长度处理,攻击者可以简单地统计通信双方的报文长度特征来确定其通信关系,从而攻破匿名保护.由上述讨论可知,Tor并不能改变其信元等长度这项特征,若将信元长度由512字节改变为其他长度,公式(2)仍然成立(将512换成对应长度即可),基于报文长度分布的识别方法依然有效.

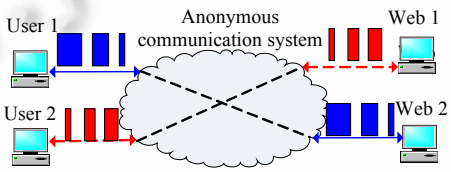


Fig.9 Application data are not set to equal size

图9 未对应用层数据进行等长度处理

依据各自的优缺点,基于TLS指纹和基于报文长度分布的识别方法可适用于不同网络监控需求.基于TLS指纹识别方法部署便捷,识别速度快,目前较适合用于需快速阻断匿名通信或及时展开匿名通信追踪的网络监管需求.基于报文长度分布的识别方法需要有前期学习过程,增加了实施复杂度,同时识别速度较慢.但该方法适合用于匿名通信内容分析[20]等网络监管需求中,因为匿名通信内容分析同样需要有学习样本和统计一定数量的报文以抽取流内容特征.当Tor改变其TLS指纹特征使得基于TLS指纹识别方法失效时,可以通过减少统计的报文数量,即通过降低识别率来使得基于报文长度分布的识别方法满足实时性的网络监控需求.

文献[21]指出:Web访问、P2P下载、FTP下载和即时通信等流量占Tor中总流量的99.92%,但其中P2P应用并不为B/S或C/S结构;同时,上传和下载数据是其典型行为特征,基于报文长度分布的识别方法可能无法识别出上层为P2P应用程序的Tor流量.然而其上传数据的流量并不经过Tor匿名通信系统转发,具体流程如图10所示.当Peer1节点需要下载某特定资源时,首先通过Tor匿名通信系统与服务器连接,下载保存该资源的其他节点信息,如图10中的Peer3.获得相应信息后,Peer1通过Tor与Peer3建立连接并下载数据.同时,Peer1节点也成为种子节点,上传相关信息至服务器.类似地,当Peer2节点需要下载同样资源时,从服务器处获得了Peer1节点信息,然后与Peer1直接建立连接并下载数据,即Peer1节点上传数据时并没有通过Tor转发,因此仍然可根据其上行流量报文长度分布特征识别出Tor流量.

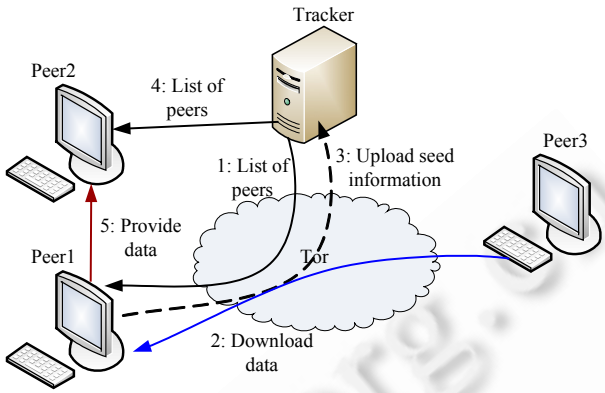


Fig.10 Download P2P resource by Tor

图 10 通过 Tor 下载 P2P 资源示意图

4 实验与分析

本文中的实验在中国教育和科研计算机网 CERNET 校园网络环境中完成.实验环境配置如图 11 所示,约有 60 台主机通过一台交换机连至 Internet 互联网.通过配置交换机的端口镜像功能将所有报文发送至一台检测主机,所有识别程序都运行于该主机上,分别测试基于 TLS 指纹和基于报文长度分布识别方法的识别率和误报率.实验中共在 10 台主机上运行 Tor 程序,访问 Web,P2P,FTP 和 IM 等不同网络资源,而其余主机上并不运行 Tor 程序,产生的流量为背景流量.实验中采用的 Tor 程序版本为 0.2.1.30.

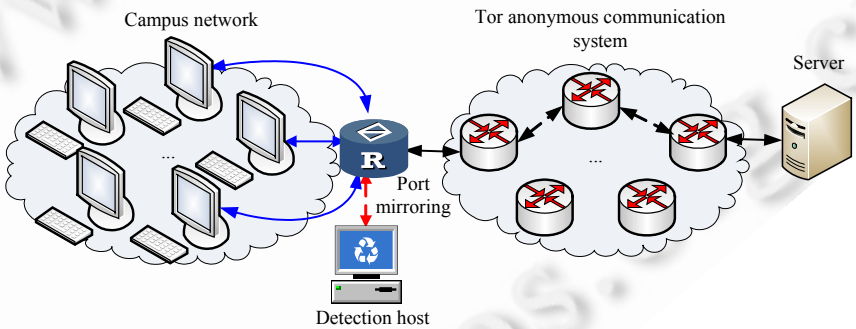


Fig.11 Experimental network setup

图 11 实验环境设置

4.1 基于TLS指纹的Tor流量识别

为保证识别率,选取的 Tor 流量 TLS 指纹特征具有稳定性,即应能识别当前不同程序版本的 Tor 流量.为分析密码套件和数字证书特征是否具有稳定性,本文首先统计出目前常用的 Tor 程序版本,然后从对应版本的源代码中分析密码套件和数字证书特征是否一致.若分析结果一致,则可认为 Tor 的 TLS 指纹特征在当前具有稳定性.

首先运行 Tor 客户端程序,客户端程序从目录服务器处下载最新的 OR 节点描述文件,然后分析该描述文件,从中提取所有 Tor 程序版本,以此作为常用 Tor 程序版本.实验时间持续为 2 周,提取的 Tor 程序版本共有 32 个.其中,最早的 Tor 程序版本为 0.2.0.26-rc,最新的稳定版本为 0.2.1.30,最新测试版本为 0.2.2.25-alpha(实验截至时间为 2011 年 5 月 8 日.实验中发现有未发布的测试程序版本,如 0.2.3.0-alpha-dev,该版本未包括于后续实验中).对此 32 个常用 Tor 程序版本对应的源代码进行分析发现,密码套件和数字证书特征均具有一致性.密码套件

在 tortls.c 文件中定义,为支持不同 OpenSSL 程序版本而采用 3 种不同密码套件.Tor TLS 证书由 Tortls.c 中 tor_tls_context_new 函数调用 tor_tls_create_certificate 函数生成.在 tor_tls_context_new 函数中生成颁发机构和拥有者名称;在 tor_tls_create_certificate 函数中赋值证书序列号为当前时间,起效时间赋值为相同时间值;证书有效时间长度定义于 or.h 该头文件中 MAX_SSL_KEY_LIFETIME 变量,时间长度为(2×60×60),即 2 小时.因此,Tor 的 TLS 指纹特征在当前具有稳定性.

为统计基于 TLS 指纹的 Tor 流量识别的识别率和误报率,本文实现了第 3.1 节中的步骤 1~步骤 6.实验中,在检测主机上运行 TLS 指纹识别程序对抓取的报文进行分析,同时记录流起始时间(第 1 个 Syn 报文的到达时间)和对应的识别结束时间,其时间差值即为识别出 Tor 流量所需的时间.实验中,在多台实验机器上首先配置 Tor 程序使用 Bridge 节点,通过 Bridge 节点下载所有 OR 节点信息,然后通过配置 UseBridges 和 UseEntryGuards 为 0 值,使得 Tor 程序与尽可能多的 OR 节点进行连接.实验持续 2 周时间,共捕获 1 357 条 Tor 连接流.将识别为 Tor 流量但目的 IP 地址并不在公开的 Tor OR 节点列表中的流量判定为误报流量,误报率计算为误报的流数量除以检测的流总数量.需要注意的是,由于一些 Tor 流量其目标 IP 地址并不属于公开的 Onion Router 列表,因而实际的误报率应小于本文实验部分得到的结果.实验结果表明,基于 TLS 指纹能 100%识别出 Tor 流量,误报率低于 0.01%,识别出 Tor 流量平均所需时间为 0.6s,识别速度快.

上述实验结果表明,基于 TLS 指纹的方法能够高效、准确地识别出 Tor 流量,但 Tor 程序可通过随机生成证书时间和颁发机构、拥有者名称等方式来抵御该识别方法.因而在实际应用中需关注 Tor 的最新版本,根据其对 TLS 密码套件和数字证书的改变而做出相应修改.

4.2 基于报文长度分布的Tor流量识别

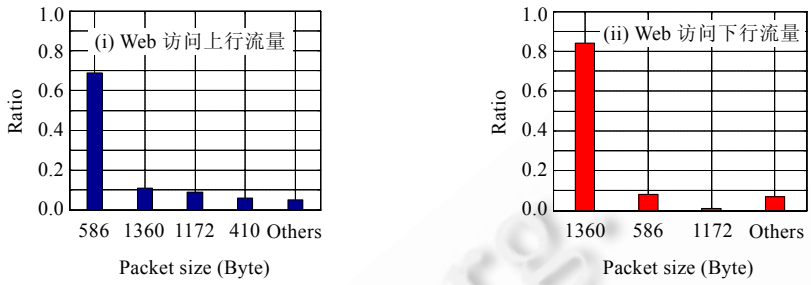
4.2.1 数据采集与分析

文献[21]指出,Web(包括 HTTP 和 HTTPS)访问、P2P 下载、FTP 下载和即时通信等流量占 Tor 总流量的 99.92%.为收集 Tor 流量的学习样本,在实验中配置浏览器、P2P 下载软件(BitTorrent(http://www.bittorrent.com/))、FTP 客户端和聊天工具等软件通过 Tor 进行网络资源访问,共收集约 10G 比特大小的 Tor 流量数据集.其他类型流量数据集采用 CAIDA Equinix-Chicago 数据集[22],支持向量机分类算法由 libsvm[23]工具实现.Equinix-Chicago 数据集由连接美国芝加哥伊利诺斯和华盛顿西雅图的主干网数据组成,含有多种不同类型的网络流量(http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA),适合作为其他类型流量的学习样本.该数据集中可能含有极少数 Tor 流量,但可将其作为样本噪声处理,而支持向量机具有一定的鲁棒性,可忽略其对实验结果的影响.

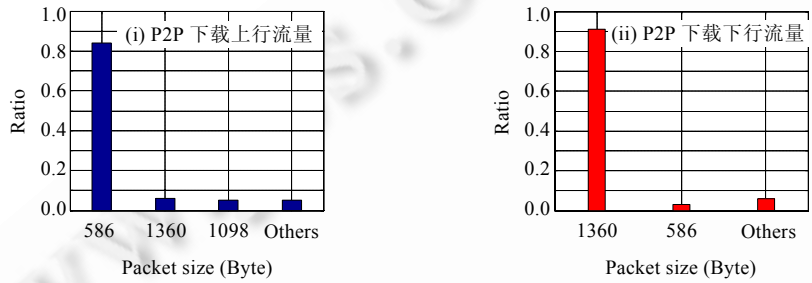
图 12 为 Tor 上行和下行流量中报文长度分布图.按报文出现的频率由高到低排序,且出现的频率大于等于 1%.Others 项表示余下的其他所有报文长度分布比例之和.实验中,最大报文段 M 值为 1 360 字节.图 12(a)为 Web 访问时的 Tor 流量中报文长度分布图,图 12(b)为 P2P 下载时的报文长度分布图,图 12(c)为 FTP 下载时的报文长度分布图,图 12(d)为即时通信(Tencent QQ(http://www.qq.com/))的报文长度分布图.各子图中的(i)和(ii)分别对应上行流量和下行流量的报文长度分布.在 Web,FTP 和 IM 等应用中,上行流量主要由客户端发出的请求命令构成,其流量均较少,因而绝大部分报文长度为少数信元长度的组合,图 12(i)中各自对应的上行流报文长度分布均符合此特征.访问 Web 页面时,浏览器会同时发出多个 HTTP 请求,从而产生 1 360 字节大小的报文.FTP 的请求则是由用户鼠标点击产生,两次请求间隔较大,不易发生排队,因而 99%的报文长度为 586 字节.即时通信产生的 1 360 字节主要是由 Tencent QQ 启动时与其服务器进行多次交互产生,当进行在线信息交互时报文长度均为 586 字节.通过 Tor 进行 P2P 下载时,其上传的数据并不经过 Tor 节点转发,因而其上行流量中报文长度分布特征与其他应用类似.数据分析结果与第 2.3 节中的结论相一致.

分析图 12(ii)得知:下行流量中报文长度分布与其他应用报文长度分布[18]相类似,主要由网络最大报文段报文和少量其他切片报文构成;并且,不同应用对应的下行流量报文长度分布差别较大(如比较 FTP 下载和即时通信),因而不能够使用其报文长度分布来检测 Tor 流量.但不同应用程序的上行流量反映了 Tor 流量的报文分布特征,主要由一个或两至三个信元组合而成,因此,本文以{586,1172,1098,410}字节为 Tor 流量的特征报文长

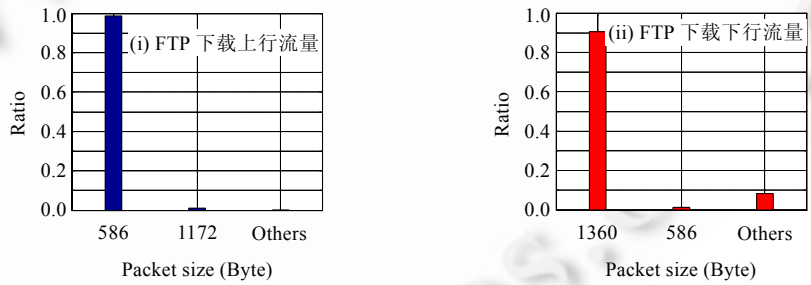
度,对应的报文结构如图 8 所示.在线识别时,可分别统计待识别流的上行流量和下行流量中报文长度分布,若其中有一条流量符合 Tor 报文长度分布特征,则判断相应的另一条流量同样为 Tor 流量.



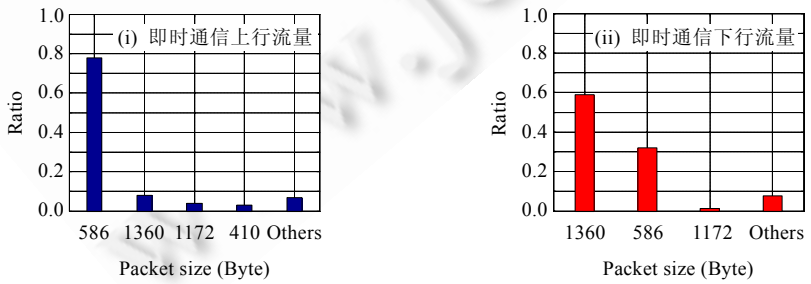
(a) 通过 Tor 访问 Web 页面时的报文长度分布



(b) 通过 Tor 下载 P2P 资源时的报文长度分布



(c) 通过 Tor 下载 FTP 资源时的报文长度分布



(d) 通过 Tor 的即时消息报文长度分布

Fig.12 Packet size distribution of Tor traffic

图 12 Tor 流量报文长度分布图

4.2.2 离线和在线实验

为达到在线识别 Tor 流量的目标,对于待识别的网络流,基于报文长度分布的识别方法应尽可能减少需要统计的报文数量.需要统计的报文数量越少,越能及时识别出 Tor 流量,同时降低对系统资源的消耗.但如果统计的报文数量太少,则可能未真实反映出流中特征报文长度分布特征,从而影响识别效果.为了得出统计不同数量的报文对识别结果的影响,本文针对 Tor 上行流量数据集和 Equinix-Chicago 数据集,分别统计流中前 25,50,100,150 和 200 个长度大于 0 的报文中特征报文的长度分布特征,并对各自的统计结果取 2/3 作为支持向量机学习样本,余下的 1/3 作为测试样本.对于 Equinix-Chicago 数据集,统计时只针对 TCP 流量且流中报文数量大于 10,这是因为目前 Tor 程序只支持 TCP 通信协议,且成功建立起一条匿名转发路径需要发送的报文数量超过 10.

计算识别率和误报率为

识别率(true positive)=(Tor 测试样本中,识别出的 Tor 流量数量)/Tor 测试样本中流总数量,

误报率(false positive)=(Equinix-Chicago 测试样本中,误判为 Tor 流量的流数量)/

Equinix-Chicago 测试样本中流总数量.

实验结果如图 13 所示,x 轴为统计的不同报文数量,左边 y 轴对应识别率,右边的 y 轴为误报率.图 13 表明,随着报文数量的增加,识别率增加,同时误报率降低.这是因为分析的报文数量越多,越能真实反映出 Tor 流量和其他类型流量的特征报文分布,从而增加识别率和减少误报.在报文数量≥50 时,识别率增加和误报率下降的趋势都不明显,因而在线识别时可以选择分析流中前 50 个报文,加快识别速度.图 13 同时表明,基于报文长度分布的识别方法具有低的误报率(10^{-5}).

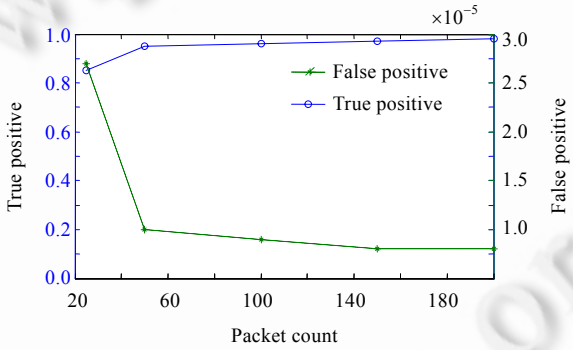


Fig.13 True and false positive for different packet count

图 13 不同报文数量对应的识别率和误报率

本文还实现了基于报文长度分布的 Tor 流量在线识别.在如图 11 所示的检测主机上,针对每条 TCP 流量,分析其前 50 个长度大于 0 的报文.以 FIN,RST 报文或时间超时标记流结束,超时时间设定为 3 分.若流中长度大于 0 的报文数量未能达到 50,则判断报文数量是否大于 10:若小于 10,则直接判断该流为其他类型流量;否则,继续识别.统计特征报文长度分布,若所有特征报文均未出现,则判断该流为其他类型流量;否则,将特征报文分布交由支持向量机判别函数判断是否为 Tor 流量.使用的判别函数为图 13 对应的实验中报文数量为 50 时形成的判别函数.同样地,记录流起始时间和识别结束时间,求出识别 Tor 流量所需的时间.

实验中,多个测试用户配置一个或多个上层应用程序使用 Tor 代理访问相应的网络资源,实验持续时间为 1 个月.实验结果显示识别率为 91%,误报率为 1.2×10^{-5}.对比离线实验结果(如图 13 所示),误报率相近,但识别率略有降低(离线实验中为 95%).这是因为 Tor 程序成功建立转发链路后,由于网络异常或 OR 节点下线等原因,该转发链路并未用于实际的数据转发,统计的报文数量小于 50.同时,运行多个上层应用程序对 Tor 流量的识别带来一定干扰,因而识别率降低.实验结果表明,识别出 Tor 流量平均所需时间为 43s,与基于 TLS 指纹的识别方法(平均为 0.6s)相比,其识别速度较慢.这是因为统计完全 50 个长度大于 0 的报文需要等待一定时间,并且实际上某

些转发链路并未用于数据转发,直至时间超时才能识别出其为 Tor 流量,增加了识别所需时间.在实际应用中,无需识别出这些未用于数据转发的 Tor 链路,因而可以通过设置较小的时间超时参数来加快识别速度.

5 结束语

本文针对匿名通信技术滥用问题,提出两种 Tor 流量识别方法.基于 TLS 指纹的识别方法将{密码套件,数字证书}作为抽取 Tor 流量 TLS 指纹特征的依据,能够高效、准确地识别出 Tor 流量,适用于需快速阻断匿名通信或及时展开匿名通信追踪的网络监管需求.但当 Tor 程序改变了其密码套件或数字证书特征时,该方法需同步做出相应的修改.基于报文长度分布的识别方法以支持向量机分类算法为核心,识别时仅需要统计网络流量中报文长度分布特征,具有高识别率和低误报率,适用于所有 Tor 程序版本.与基于 TLS 指纹的识别方法相比需要有前期学习过程,并且实时性较差,可用于匿名通信内容分析等网络监管需求中.

本文提出的识别方法均为报文级的,下一步的研究包括:研究基于流和主机行为特征的 Tor 流量识别.此外,如何突破现有匿名通信系统设计的固有理念(如信元长度相等),使得匿名通信流量与其他网络流量报文长度分布的差别变小,是另一项具有挑战性的研究课题.

References:

[1] Moller U, Cottrell L, Palfrader P, Sassaman L. Mixmaster protocol-V2. IETF Internet draft, 2003. http://tools.ietf.org/html/draft-sassaman-mixmaster-03

[2] Danezis G, Dingledine R, Mathewson N. Mixminion: Design of a type III anonymous remailer protocol. In: Proc. of the 2003 IEEE Symp. on Security and Privacy. 2003. 2−15. [doi: 10.1109/SECPRI.2003.1199323]

[3] Dingledine R, Mathewson N, Syverson P. Tor: The second-generation onion router. In: Proc. of the 13th USENIX Security Symp. 2004. http://www.torproject.org/tor-design.pdf

[4] Berthold, O, Federrath, H, Kopsell, S. Web Mixes: A system for anonymous and unobservable Internet access. In: Proc. of the Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability. LNCS 2009, Berlin: Springer-Verlag, 2000. 115−129. [doi: 10.1007/3-540-44702-4_7]

[5] Chaum DL. Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM, 1981,24(2): 84−88. [doi: 10.1145/358549.358563]

[6] JAP Team. Mixes for privacy and anonymity in the Internet documentation. 2006. http://anon.inf.tu-dresden.de/develop/doc/mix_short/index.html

[7] Wang XY, Chen S, Jajodia S. Tracking anonymous peer-to-peer VoIP calls on the Internet. In: Proc. of the 12th ACM Conf. on Computer and Communications Security. Alexandra: ACM Press, 2005. 81−91. [doi: 10.1145/1102120.1102133]

[8] Fu XW, Zhu Y, Graham B, Bettati R, Zhao W. On flow marking attacks in wireless anonymous communication networks. In: Proc. of the IEEE Int'l Conf. on Distributed Computing Systems. Columbus: IEEE Press, 2005. 493−503. [doi: 10.1109/ICDCS.2005.55]

[9] Ling Z, Luo JZ, Yu W, Fu XW, Xuan D, Jia WJ. A new cell counter based attack against Tor. In: Proc. of the 16th ACM Conf. on Computer and Communications Security (CCS). 2009. 578−589. [doi: 10.1145/1653662.1653732]

[10] Wang XG, Luo JZ, Yang M. A double interval centroid-based watermark for network flow traceback. In: Proc. of the 14th Int'l Conf. on Computer Supported Cooperative Work in Design. 2010. 146−151. [doi: 10.1109/CSCWD.2010.5471985]

[11] Houmansadr A, Borisov N. SWIRL: A scalable watermark to detect correlated network flows. In: Proc. of the 18th Annual Network & Distributed System Security Symp. (NDSS). 2011.

[12] Dingledine R, Mathewson R. Design of a blocking-resistant anonymity system. 2008. https://svn.torproject.org/svn/tor/tags/tor-0_1_2_10_rc/doc/design-paper/blocking.pdf

[13] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks. IEEE/ACM Trans. on Networking, 2004,12(2):219−232. [doi: 10.1109/TNET.2004.826277]

[14] Sen S, Spatscheck O, Wang DM. Accurate, scalable in-network identification of P2P traffic using application signatures. In: Proc. of the 13th Int'l Conf. on World Wide Web. New York: ACM Press, 2004. 512−521. [doi: 10.1145/988672.988742]

[15] Constantinou F, Mavrommatis P. Identifying known and unknown peer-to-peer traffic. In: Proc. of the 5th IEEE Int'l Symp. on Network Computing and Applications. 2006. 93−102. [doi: 10.1109/NCA.2006.34]

[16] Silveira F, Diot C, Taft N, Govindan R. ASTUTE: Detecting a different class of traffic anomalies. In: Proc. of the ACM SIGCOMM 2010 Conf. on SIGCOMM. 2010. 267−278. [doi: 10.1145/1851182.1851215]

[17] Kanda Y, Fukuda K, Sugawara T. A flow analysis for mining traffic anomalies. In: Proc. of the IEEE Int'l Conf. on Communications. 2010. [doi: 10.1109/ICC.2010.5502463]

[18] Sinha R, Papadopoulos C, Heidemann J. Internet packet size distributions: Some observations. Technical Report, ISI-TR-2007-643, USC/Information Sciences Institute, 2007.

[19] Zhang XG. Introduce to statistical learning theory and support vector machines. ACTA Automatic Sinica, 2000,26(1):32−42 (in Chinese with English abstract).

[20] Herrmann D, Wendolsky R, Federrath H. Website fingerprinting: Attacking popular privacy enhancing technologies with the multinomial naive-Bayes classifier. In: Proc. of the ACM Workshop on Cloud Computing Security. 2009. [doi: 10.1145/1655008.1655013]

[21] McCoy D, Bauer K, Grunwald D, Kohno T, Sicker D. Shining light in dark places: Understanding the Tor network. In: Proc. of the 8th Int'l Symp. on Privacy Enhancing Technologies. 2008. 63−76. [doi: 10.1007/978-3-540-70630-4_5]

[22] Walsworth C, Aben E, Claffy KC, *et al*. The CAIDA anonymized 2010 Internet traces. 2010. http://www.caida.org/data/passive/passive_2010_dataset.xml

[23] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm

附中文参考文献:

[19] 张学工.关于统计学习理论与支持向量机.自动化学报,2000,26(1):32−42.

何高峰(1984－),男,安徽安庆人,博士生,主要研究领域为网络安全,匿名通信.
E-mail: hegaofeng@seu.edu.cn

罗军舟(1960－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为下一代网络体系结构,网络安全与管理,网格与云计算,无线局域网.
E-mail: jluo@seu.edu.cn

杨明(1979－),男,博士,副教授,CCF 会员,主要研究领域为网络安全.
E-mail: yangming2002@seu.edu.cn

张璐(1983－),男,博士生,主要研究领域为网络安全.
E-mail: luzhang@seu.edu.cn