

抗扰动的网站指纹鲁棒识别方法

张静茜^{1,2}, 李腾耀^{1,2}, 涂宇宽^{1,2}, 罗向阳^{1,2}

(1.信息工程大学, 河南 郑州 450001; 2.网络空间态势感知河南省重点实验室, 河南 郑州 450001)

摘要: 网站指纹识别通常基于用户浏览网页的流量中暴露的网站指纹特征识别其访问的目标网站, 此类技术对于刻画用户的匿名访问行为、提升以Tor为代表的匿名流量监管治理具有重要意义。然而, 大量针对性的防御方法通过算法对具有区分性的关键流量特征进行扰动, 造成现有网站指纹识别方法的识别准确率大幅下降。现有鲁棒性最优的网站指纹识别(robust fingerprinting, RF)方法能够在多种防御方法上保持较好的性能, 但难以抵抗针对性防御方法——RF Countermeasure。为此, 提出了基于混合特征矩阵的抗扰动网站指纹鲁棒识别方法。区别于RF, 该方法使用包累积长度代替累积包数量作为数据包级特征; 运用信息泄露理论分析流特征鲁棒性, 基于数据包方向分布和持续同方向数据包数量, 构建会话级的鲁棒流特征; 综合数据包级和会话级的特征, 构建抗多种防御扰动的混合特征矩阵, 以该矩阵为输入, 采用深度网络模型对网站指纹进行分类识别。基于深度指纹(deep fingerprinting, DF)公开数据集开展了大量实验, 结果表明, 提出的方法在防御方法RF Countermeasure上的准确率达到95.4%, 与现有RF方法相比提高了21.2%, 同时该方法在其他典型防御场景下也保持了良好的识别性能。

关键词: 网站指纹识别; 匿名流量分类; 混合特征矩阵; 鲁棒指纹; 流量分析

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024086

Robust identification method of website fingerprinting against disturbance

ZHANG Jingxi^{1,2}, LI Tengyao^{1,2}, TU Yukuan^{1,2}, LUO Xiangyang^{1,2}

1. Information Engineering University, Zhengzhou 450001, China

2. Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou 450001, China

Abstract: Website fingerprinting usually identifies the target website visited by users based on the website fingerprint characteristics exposed in the web traffic. It is essential in tracking users' anonymous access behaviors and

收稿日期: 2024-05-18; 修回日期: 2024-06-24

通信作者: 李腾耀, litengyao@aliyun.com

基金项目: 国家重点研发计划(2022YFB3102900); 国家自然科学基金(62172435, U23A20305); 河南省重点研发专项(221111321200)

Foundation Items: The National Key Research and Development Program of China (2022YFB3102900), The National Natural Science Foundation of China (62172435, U23A20305), Key Research and Development Project of Henan Province (221111321200)

引用格式: 张静茜, 李腾耀, 涂宇宽, 等. 抗扰动的网站指纹鲁棒识别方法[J]. 网络与信息安全学报, 2024, 10(6): 137-150.

Citation Format: ZHANG J X, LI T Y, TU Y K, et al. Robust identification method of website fingerprinting against disturbance[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 137-150.

improving the anonymous traffic governance, especially on Tor network flows. However, many defense mechanisms emerged to disturb the distinctive traffic patterns, which results in website fingerprint identification failure. The existing website fingerprint identification method with the best robustness named RF can maintain good performance against various defense methods, but it is difficult to resist the targeted defense method RF Countermeasure. An anti-defense website fingerprinting based on hybrid feature matrix (ADF) was proposed. Unlike RF, ADF used the cumulative packet length instead of the cumulative packet number as the packet-level feature. On the basis of analyzing information leakage value of flow features, ADF constructed the robust flow features of the session level using packet direction distribution and the number of continuous packets in the same direction. Subsequently, a hybrid feature matrix (HFM) was constructed to resist various defense disturbance by combining the features of both packet-level and session-level. With the matrix as input, a robust flow classifier with convolutional neural network was established. Through extensive experimental analysis on the dataset provided by DF, the accuracy under RF Countermeasure is 95.4%, which is 21.2% higher than RF. This method also maintains good identification performance under other state-of-the-art defenses.

Keywords: website fingerprinting identification, anonymous traffic classification, hybrid feature matrix, robust fingerprint, traffic analysis

0 引言

在以 Tor 为代表的匿名通信协议支持下, 用户访问网站的会话流量往往具有显著的双向匿名特性, 难以直接通过数据包包头及其数据载荷有效解析用户与网站之间的交互关系, 使潜在的非访问行为极易脱离网络监管, 成为网络空间安全治理的死角。然而, 由于不同网站在服务内容、响应模式等方面的差异性, 网络流量的包时间间隔、包长度、包方向等流特征会出现与网站类别相耦合的变化特性, 形成不同网站访问过程中会话流量的指纹信息。基于这些网站的流量指纹信息, 能够在庞杂的网站流量中解析用户和网站的对应访问关系。当前, 网站指纹识别方法已成为破解匿名通信流量监管难题的重要手段^[1]。

为了实施网站指纹识别, 网站指纹识别方一般位于客户端和 Tor 入口节点之间, 捕获网站流量并提取其侧信道信息, 并据此进行网站类别的分析与识别, 如图 1 所示。

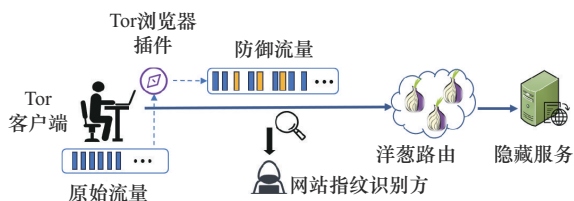


图 1 网站指纹运用场景

Figure 1 Scenario of website fingerprint identification

一般地, 网站指纹识别方仅具有嗅探流量的能力, 不能对流量进行延迟、丢弃和修改。由于网站指纹防御可通过 Tor 浏览器插件进行部署, 网站指纹识别方嗅探到的流量可能是经防御方法混淆后的流量, 防御方法的目标可建模表示为:

$$\max_{\delta} \text{Loss}(\text{Output}(f + \delta), \text{Output}(f)) \quad (1)$$

其中, f 表示原始流量的特征序列, δ 表示网站指纹防御方法带来的扰动, $\text{Output}(\cdot)$ 表示分类器的输出, $\text{Loss}(\cdot)$ 表示损失函数, 一般使用交叉熵损失函数。

网站指纹防御方通过引入扰动 δ 使模型输出结果与无防御条件下模型输出结果的差异程度最大。在该场景下, 需要挖掘建模足够鲁棒的网站指纹特征, 对抗网站防御引入的随机化扰动, 并使:

$$\min \left(\max_{\delta} \text{Loss}(\text{Output}(\hat{f} + \delta), \text{Output}(\hat{f})) \right) \quad (2)$$

其中, $\hat{f}$ 表示原始流量的鲁棒特征序列。在网站指纹防御扰动条件下, 依托该特征仍能够最小化模型的分类损失, 逼近正确的分类结果。

现有的网站指纹识别方法, 经历了从人工遴选统计特征到机器自主提取特征的发展历程, 深度指纹 (deep fingerprinting, DF) 方法^[2], 构建了以卷积块为基础单元的分类模型, 依托包方向特征序列作为输入, 在无防御条件下实现了网站流量的有效分类。随后, Tik-Tok^[3]、Var-

CNN^[4]等网站指纹识别方法相继被提出,将包方向和包时间等多样化特征进一步引入网站指纹识别方法中,网站流量的分类准确率提升到98.4%。

然而,随着网站指纹识别的攻防对抗,涌现出了包含WTF-PAD^[5]、Front^[6]、RF Countermeasure^[7]等网站指纹防御方法。这些方法通过插入虚拟包或延迟数据包的方式,向网络流量中注入干扰,使原会话流量的流特征序列在特征空间出现显著偏移,网站指纹识别的准确率下降至80%以下。基于Pluggable Transports插件机制,出现了网站指纹防御的浏览器插件,配合Tor浏览器一同使用,能够实现网站指纹防御方法的实网部署。因此,网站指纹防御对网站指纹识别的影响,已经成为网站指纹识别无法绕过的现实挑战^[8]。

为避免网站指纹防御的影响,需提取鲁棒的网站流量特征。Shen等^[7]构建了鲁棒指纹识别(robust fingerprinting, RF)方法,将指定时间窗口内的包数量作为鲁棒特征,显著提升了网站指纹识别的鲁棒性。但是,RF Countermeasure防御方法在获取到网站流量中对识别分类贡献显著的富信息区域后,通过向这些富信息区域引入虚拟包插入和数据包延迟的针对性干扰,使RF的分类准确率下降至52.59%。因此,如何在RF Countermeasure防御条件下,保持对网站指纹的识别性能是一个值得研究的问题。

为此,本文通过挖掘能够对抗虚拟包插入、数据包延迟等典型防御干扰操作的鲁棒流特征,构建混合特征矩阵(hybrid feature matrix, HFM),提出抗多样化防御的网站指纹(anti-defense website fingerprinting, ADF)识别方法,主要工作如下:

1) 构建更加鲁棒的混合特征矩阵**HFM**。融合数据包包方向、给定时间窗口内包累积长度、持续同方向数据包的出现次数3类流特征,深度挖掘网络流量中包方向、包长度、包数量携带的网站类别信息,兼顾局部特征和全局特征的优势,构建了混合特征矩阵,提升了抗多样化网站指纹防御的能力。

2) 提出ADF识别方法。基于RF深度学习模

型结构,以混合特征矩阵为输入,改进提出了能够有效对抗RF Countermeasure的鲁棒网站指纹识别方法,克服了第三方在富信息域注入干扰导致分类性能大幅下滑的挑战。

3) 进行多样化防御场景下的性能测试。针对DF采集的数据集,在封闭世界和开放世界两种场景下,对比分析了ADF与DF、TikTok、RF的分类性能,构造了不同数据包填充长度条件下的网站指纹防御策略,并在WTF-PAD、Front、RF Countermeasure等防御方法下验证了ADF的鲁棒性,在RF Countermeasure防御条件下ADF的准确率达到95.4%,其鲁棒性优于现有方法。

1 相关工作

网站流量中的指纹信息,往往使用机器学习进行提取和分类。早期网站流量指纹识别方法,主要依托包括支持向量机、随机森林等机器学习模型在内的方法进行网站分类,典型方法有 k -NN^[9]、CUMUL^[10]和 k -FP^[11]等,这类方法一般通过人工设计遴选流量特征的方式构建网站流量的指纹信息,并以此作为模型输入进行有监督地分类识别,这些方法在无防御、大规模训练数据、有限网站数量条件下能够准确识别网站类别。但是,这类模型的准确性高度依赖于人为构造特征的有效性,无法抵抗WTF-PAD等典型防御方法对流量的针对性扰动。

由于人为构造特征的局限性,深度学习逐步引入用于提取更深层的特征^[12-13]。为了进一步提升网站指纹识别的准确率,避免多种防御干扰引入的分类偏差,涌现出了包括卷积神经网络^[2-3](convolutional neural network, CNN)、三重网络^[14]、生成对抗网络^[15]等在内的深度学习指纹识别模型。这些深度学习模型能够自动学习类别间具有区分性的特征,在大幅简化人工设计特征流程的同时,显著提升分类准确率,特别是在有防御、少样本、大规模网站类别条件下实现了性能的提升优化。Sirinam等^[2]提出DF方法,使用包方向序列作为输入,并利用具有复杂架构设计的卷积神经网络。在无防御条件下准确率超过98%,优于之前所有的网站指纹识别方法。同时,DF首次击败了防御方法WTF-PAD。但对于防御方法Front和Walkie-

Talkie^[16], 准确率下降到 80% 以下。为了充分利用多种流特征挖掘流量的身份信息, Rahman 等^[3]研究了时间信息促进分类器识别站点的程度, 提出 Tik-Tok 方法。相较 DF 等只使用包方向单一特征的网站指纹识别方法, Tik-Tok 综合考虑了流量定时突发性的影响, 采用包方向和包时间构建流量融合指纹, 进一步扩展了网站指纹分类的适用场景。

然而, 在匿名网络流量中可采集的可靠样本相对较少, 使在少量可靠样本上进行网站指纹模型的训练面临性能瓶颈。一方面, 目标网站由于在线时长不确定、隐藏服务访问少等, 使这类网站的样本流量相对较少。为此 Sirinam 等^[14]提出 TF (triplet fingerprinting), 使用三重网络进行少样本学习, 减少了数据收集和训练的工作量, 并能适应不同网络条件的变化。Bhat 等^[4]提出 Var-CNN 方法, 基于残差网络, 使用包方向、包时间间隔和元数据作为输入, 构建了一个集成模型, 仅需要较小的样本进行训练, 即可获得与 DF 相当的分类性能。另一方面, 目标网站的流量样本易受到网站指纹防御技术的干扰, 使采集的流量样本在包方向、包时间间隔、包长度等维度出现不确定偏移, 造成网站指纹识别的错误分类。为解决网站指纹防御引入的流量扰动, Shen 等^[7]通过生成流量聚合矩阵 (traffic aggregation matrix, TAM), 刻画了网站流量泄露的身份信息, 构建了适应多种网站指纹防御扰动的鲁棒网站指纹特征。在 TAM 中, 每个数值代表特定的时间范围内当前流量上行包/下行包的数量。利用 TAM 训练基于 CNN 的分类器, 能在不同防御扰动的网站流量中学习通用的网站识别特征。在有防御条件下, RF 的平均分类准确率比网站指纹方法 Tik-Tok 提高了 8.9%, 是当前先进的鲁棒识别方法。表 1 给出了 DF、Tik-Tok 和 RF 准确率对比, 结果验证了 RF 在无防御和有防御的条件下性能的优越性和鲁棒性。

但是 Shen 等^[7]在提出 RF 的同时, 也提供了针对性防御方法——RF Countermeasure。该方法通过类激活映射 (class activation mapping, CAM) 提取一组历史 TAM 的富信息区域, 使用

虚拟包插入和数据包延迟的策略让某网站的 TAM 表现为另一不同网站的富信息区域的集合, 从而有效混淆网站的类别特征。该防御方法成功使 RF 的分类准确率从无防御场景的 98.83% 下降至 52.59%。

表 1 DF、Tik-Tok 和 RF 准确率对比

Table 1 Comparison of DF, Tik-Tok and RF accuracy^[7]

方法	准确率			
	无防御	WTF-PAD	Front	RF Countermeasure
DF	98.40%	90.85%	76.85%	—
Tik-Tok	98.45%	93.80%	84.79%	—
RF	98.83%	96.58%	93.34%	52.59%

因此, 现有方法在有防御条件下实施网站指纹识别仍面临防御多样化和扰动随机化的现实挑战, 在无防御、可靠样本较少的条件下实现网站流量类别的精准分类成为当前亟待解决的技术难题。

2 ADF 方法

本文依托数据包的多维流特征, 构建包含包方向、包累积长度、持续同方向数据包的出现次数在内的混合特征矩阵, 基于 RF 模型提出 ADF 识别方法, 适应 WTF-PAD、Front、RF Countermeasure 等在内的主流网站防御方法的针对性干扰, 实现有防御、有干扰条件下网站类别的有效划分。

ADF 方法主要包括两个部分, 一是通过构建 HFM 挖掘鲁棒网站指纹特征, 构建流量身份的有效表示, 在有防御条件下有效建模网站的身份属性信息; 二是通过构建 CNN 分类器模型, 基于 HFM 进行网站流量的准确分类, 对流量所属的网站类别进行划分。网站指纹识别方在客户端和入口节点之间嗅探流量, 提取特征并构建 HFM, 据此训练一个有监督的分类器, 判别当前用户正在访问的目标网站。ADF 方法总体框架如图 2 所示。

2.1 混合特征矩阵构建

为充分利用流特征携带的身份信息, HFM 结构上设计为一个 $4 \times n$ 的矩阵, 前 2 行表示流量的局部特征, 后 2 行表示流量的全局特征。第 1 行表示上行 (即客户端发送) 流量在大小为

s 时间窗口内的累积包长度，第2行表示下行（即客户端接收）流量在大小为 s 时间窗口内的累积包长度；第3行表示上行流量在大小为 $N \times s$ 时间窗口内的 burst 数量，第4行表示下行流量在大小为 $N \times s$ 时间窗口内的 burst 数量。burst 是同方向上一组连续的包序列，同一 burst 内部所有数据包的方向相同。

2.1.1 局部特征提取

划分时间窗口。传统的方法如 DF、Tik-Tok 等，使用单个包级别的特征。这样细粒度的特征对网站指纹的贡献较高，在无防御的条件下能获得非常好的分类性能，但针对性的防御方法也能轻易带来扰动，鲁棒性较差。防御方法有两种典型的混淆策略：一是延迟正常数据包的发送；二是额外填充虚拟包。这两种策略直接影响包时间序列和包交互关系，从而对单个包级别的特征产生重大影响。基于上述问题，ADF 对流量进行时间窗口的划分，使用一定长度 s 时间窗口内的统计特征，以此提高特征的鲁棒性。因为对于包延迟，在保证用户体验的约束下，延迟后的包大概率仍落在时间窗口内。而对于包填充，虚拟包的总数能够分散到多个时间窗口，不会造成剧烈的扰动。

引入包累积长度。由于网站提供的服务和开

发方式各异，不同网站的交互模式和数据体量存在差异，网站指纹是否能有效刻画差异信息，是能否准确识别网站类别的关键之一。与以往的网站指纹方法一致，ADF 采用包方向刻画网站的交互模式。数据体量可以使用包数量和包长度进行表征，选择包长度的原因是其单位为字节，相比包数量的刻画粒度更细，并且与时间窗口相结合后，集中填充超短垃圾包的防御策略会对累积包数量造成剧烈扰动，但对累积包长度的影响较小。从这个角度看，累积包长度具有更强的鲁棒性。

基于上述设计思想，具体的局部特征提取方法为：每条流的最大加载时间记为 T ，用时间戳和长度标识数据包 $p_n = \langle t_n, l_n \rangle$ ，则矩阵第1、2行元素计算方法为：

$$\text{HFM}[0][i] = \sum_{n=1}^{k_0} l_n \quad (3)$$

$$\text{HFM}[1][i] = \sum_{n=1}^{k_1} l_n \quad (4)$$

其中， $i (i \in [0, \lceil \frac{T}{s} \rceil])$ 是矩阵列索引，代表时间窗口的序号， k_0 是第 i 个时间窗口内上行包数量， k_1 是第 i 个时间窗口内下行包数量。

为了验证局部特征的有效性，本文使用网站

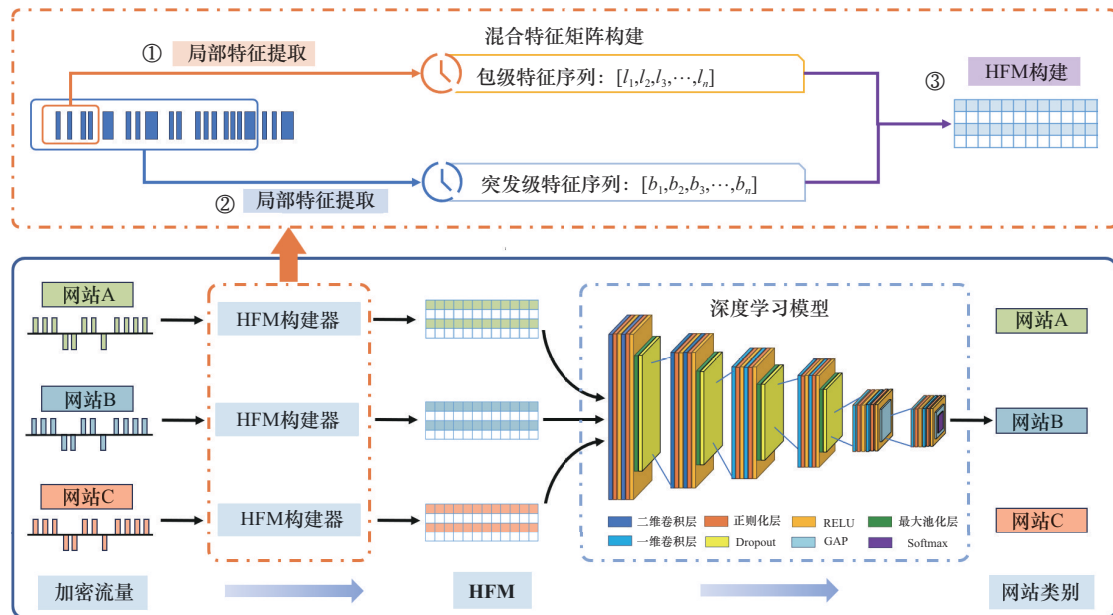


图2 ADF方法总体框架

Figure 2 The overall framework of ADF

指纹信息泄露的度量理论^[17]来量化可以从特征中获取到的信息量。封闭世界中的信息泄露值 $I(F; W)$ 定义为:

$$I(F; W) = H(W) - H(W|F) \quad (5)$$

其中, W 为被监控的网站, F 为特定表示的特征, $H(\cdot)$ 为熵。

本文选取 DF (包方向)、Tik-Tok (包方向×包时间)、RF (包数量/秒) 和 ADF 的特征分别在无防御数据集和 Front 数据集上进行信息泄露分析, 每个网站随机选择 100 条流, 共 95 个网站, 各特征的信息泄露分析结果如图 3 所示, 各特征信息泄露平均值见表 2。

从图 3(a)-(d) 和表 2 第 1~4 个特征的信息泄露分析可知, 本文提取的一定时间窗口内累积包长度的局部特征, 在原始流量上的信息泄露值最高, 最大值达到 1.86 bit, 平均值为 1.39 bit, 高于 DF、Tik-Tok 和 RF 特征的平均值 0.51 bit、0.73 bit 和 1.29 bit, 充分说明局部特征的有效性。

在 RF Countermeasure 防御下, 本文所提特征 4 的信息泄露值为 1.44 bit, 略低于 RF 特征的 1.45 bit。然而, RF Countermeasure 防御机制与

Front 和 WTF-PAD 不同, 该方法通过模仿其他类别网站的 TAM 富信息域来诱导分类器误分类, 使提取出的包数量/秒和包累积长/秒特征值近似于其他类别网站的富信息区域特征值。因此, 对于 RF Countermeasure 防御机制而言, 流特征的信息泄露值需要维持在一个合理区间内, 不宜过高, 导致显著的误分类, 也不宜过低, 导致分类特征丢失。尽管特征 3 和 4 的信息泄露值在 RF Countermeasure 下分别高达 1.45 bit 和 1.44 bit, 甚至高于无防御场景, 但两个特征泄露的信息是由防御方法特意构造、指向错误类别的。因此在以模仿其他网站类别机制为基础的防御场景下, 特征 4 误导分类器的程度略轻于特征 3。此外, 从无防御场景到 Front 和 WTF-PAD, 特征 4 的平均信息泄露值仅下降 0.09 bit 和 0.27 bit, 下降后信息泄露值仍居首位, 可有效支撑网站指纹的分类与识别, 这充分说明了本文所选局部特征的鲁棒性。

2.1.2 全局特征提取

虽然累积包长度很好地刻画了流量的局部特征, 但由于有防御条件下局部特征存在针对性扰动 (如 RF Countermeasure) 的可能性, 有必要引入流量的全局性信息, 作为网站指纹识别的重

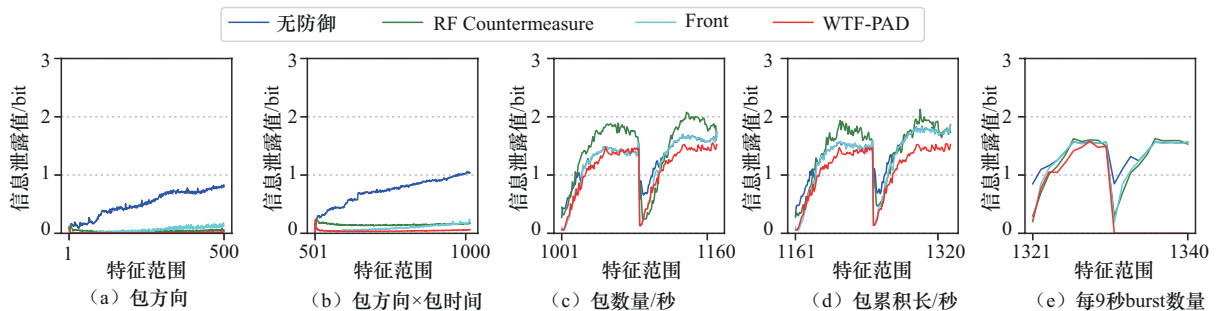


图 3 各特征的信息泄露分析结果

Figure 3 Information leakage analysis results of features

表 2 各特征信息泄露平均值

Table 2 Average information leakage value of features

序号	特征	平均值/bit			
		无防御	RF Countermeasure	Front	WTF-PAD
1	包方向(DF)	0.51	0.04	0.06	0.01
2	包方向×包时间(Tik-Tok)	0.73	0.15	0.09	0.04
3	包数量/秒(RF)	1.29	1.45	1.22	1.12
4	包累积长/秒(ADF)	1.39	1.44	1.30	1.12
5	每 9 秒 burst 数量(ADF)	1.36	1.24	1.26	0.59

要支撑。在网络流量中, **burst** 是一个重要的特征^[18]。**HFM** 在一个较大时间窗口 (整数 N 倍的 s) 内, 分别统计上行流量和下行流量的 **burst** 数量作为特征, 从而表示更为宏观的信息。包延迟的策略考虑用户体验, 一般不会切分 **burst**; 在真实防御场景下填充的垃圾包往往伴随真实数据发送, 所以在一段时间内上行和下行的 **burst** 数量基本趋于稳定。

基于上述设计思想, 具体的全局特征提取方法为: 规定每条流的最大加载时间为 T , 由于 **burst** 是一组包序列, 可能会出现跨时间窗口的现象, **ADF** 使用 **burst** 包时间均值决定当前 **burst** 归属的时间窗口。**HFM** 中第3、4行表示为:

$$\mathbf{HFM}[2][j] = \sum_{n=1}^{k_2} 1 \quad (6)$$

$$\mathbf{HFM}[3][j] = \sum_{n=1}^{k_3} 1 \quad (7)$$

其中, $j (j \in [0, \left\lfloor \frac{T}{N \times s} \right\rfloor])$ 是矩阵列索引, 代表较大时间窗口的序号, k_2 是第 i 个时间窗口内上行 **burst** 数量, k_3 是第 i 个时间窗口内下行 **burst** 数量。

从图3和表2第1~5个特征的信息泄露分析结果可知, 本文提取的较大时间窗口内 **burst** 数量的全局特征, 在原始流量上的信息泄露值仅次于局部特征, 高于 **DF**、**Tik-Tok** 和 **RF** 的平均值 0.51 bit、0.73 bit 和 1.29 bit, 充分说明全局特征的有效性。

从无防御场景到 **RF Countermeasure** 和 **Front**, 特征5的信息泄露值仅下降 0.12 bit 和 0.1 bit, 仍能够有效支撑网站指纹的分类与识别, 这充分说明了全局特征的鲁棒性。在 **WTF-PAD** 防御下, 因为 **WTF-PAD** 防御产生的数据包定长, 包累积长度是包数量的整数倍, 所以两个特征信息泄露值相等; 全局特征的信息泄露值为 0.59 bit, 较局部特征有所下降, 其原因在于 **WTF-PAD** 会在 **burst** 较大的时间间隔中填充虚假 **burst**、由数据包收发事件引入的虚拟数据包填充可能会分割当前 **burst**, 削弱 **burst** 的全局特征表达能力, 但全局特征的信息泄露仍明显高于 **DF** 和 **Tik-Tok** 方法使用的特征。尽管在 **WTF-PAD** 防御下, **RF** 使用的包数量/秒特征优于全局特征, 但由于本文所

提方法旨在对 **RF** 进行优化, 因此更加关注 **RF Countermeasure** 防御下的特征泄露值, 全局特征误导分类的程度较轻, 更符合方法需求。同时, **WTF-PAD** 对全局特征的影响, 可以用特征融合的方式减轻或消除。因此, 本文将全局特征与局部特征进行融合, 构建网站指纹识别的鲁棒流特征。

2.1.3 HFM 构建

由于累积包长度和 **burst** 数量的量纲不同, 为了确保不同特征对模型训练的影响力相对平衡, **ADF** 采用最大最小归一化将矩阵的值映射到 (0,1], 如式(8)所示。

$$x_i = \frac{x_i - \text{Min}}{\text{Max} - \text{Min}} \quad (8)$$

其中, x_i 是 **HFM** 中的数值, **Min** 是 x_i 所在行数据的最小值, **Max** 是 x_i 所在行数据的最大值。

在综合使用局部特征和全局特征的情况下, 可完成 **HFM** 的构建。为便于阐述, 给出 **HFM** 构建的一个具体示例, 如图4所示。给定一条最大加载时间为 T 的流量, **HFM** 的第1行是大小为 s 时间窗口内的上行方向的累积包长度, 对应图中紫色包; 第2行是下行方向, 对应蓝色包, 其余同第一行; 第3行是大小为 $2 \times s$ 时间窗口内上行方向的 **burst** 数量, 第4行是下行方向, 其余同第3行。

综上, **HFM** 生成过程如算法1所示。第1个阶段用于提取局部特征, 统计大小为 s 时间窗口内上下行包数量; 第2个阶段用于提取全局特征, 统计大小为 $N \times s$ 时间窗口内上下行 **burst** 数量; 最后对矩阵进行最大最小归一化, 得到涵盖流量局部特征和全局特征的混合特征矩阵。

算法1 HFM 生成

输入 会话流量 F , 时间窗口的长度 s , 时间窗口切分数量 n

输出 $\mathbf{HFM} = \{ m_{ij} | i \in [0,3], j \in [0, n-1] \}$

1) 初始化 **HFM** //矩阵元素初值设为0

2) **for each packet** $p_k = \langle t_k, d_k, l_k \rangle \in F$ **do**

3) $j \leftarrow \left\lceil \frac{t_k}{s} \right\rceil$ //计算列索引

4) **if** $j \leq n$ **then**

5) $i \leftarrow d_k < 0 ? 0 : 1$ //包方向确定行

幅为 1×1 ，填充为 1×1 ，最大池化层核大小为 2×2 ，丢弃层丢弃概率为0.1；一维卷积块3中的卷积层输出通道数为128，核大小为3，步幅为1，填充为1，最大池化层核大小为3，丢弃层丢弃概率为0.3；一维卷积块4中的卷积层输出通道数为256，其余与一维卷积块3一致；全局平均池化中的卷积层1输出通道数为512、卷积层2输出通道数为网站类别数，其余与一维卷积块3中的卷积层一致。

3 实验分析

为了验证ADF的有效性，本节基于DF数据集生成多种防御流量，依次在封闭世界和开放世界场景下进行ADF性能分析。

网站指纹识别方法通常在两种场景下进行评估：封闭世界场景和开放世界场景。训练一个能够识别用户可能访问的所有目标网站的分类器是不现实的，因此网站指纹识别方往往在一组感兴趣的有限网站集合上训练分类器，这个集合被称为监控集，其他所有网站称为非监控集。封闭世界场景假设用户只访问监控集中的网站，网站指纹识别方使用监控网站的流量训练分类器，这种假设尽管不现实，但对评估模型性能很有用；开放世界场景则更贴近现实，用户不仅会访问监控集网站，还会访问其他大量的非监控集网站，网站指纹识别方使用监控网站的流量和部分非监控网站流量训练分类器，分类器需要判断当前用户访问的网站是否属于监控集，如果属于监控集，则进一步判定具体的网站类别。

3.1 数据集

DF数据集被DF、Tik-Tok、RF等主流方法广泛使用，并被作为权威数据集进行方法性能的验证。因此，本文同样使用DF数据集作为分析基础，并且本文将基于该数据集进行方法间的对比分析。在该数据集中，封闭世界数据集包含来自Alexa排名前100站点中的95个网站，每个网站1 000条流，共95 000条流；开放世界数据集包含来自Alexa排名前50 000站点中的40 000个网站（不包括封闭世界的95个网站），每个网站1条流。在封闭世界场景中，训练集和测试集的比例为9:1，每个网站随机选择900条流用于训练，剩余100条用于测试；在开放世界场景中，

训练集包含封闭世界95个网站的900条流，加上开放世界20 000条流，共105 500条流；测试集包含封闭世界95个网站的100条流，加上开放世界20 000条流，共29 500条流。用于训练和测试的数据没有交叉。

此外，本文在DF数据集的基础上，根据WTF-PAD、Front和RF Countermeasure的源代码，生成对应的防御流量，扩展形成包含多样化防御干扰的综合数据集，防御参数设置如表3所示。

表3 防御参数设置

Table 3 Defense hyperparameter settings

序号	防御	参数
1	WTF-PAD ^[5]	Config = normal_rcv
2	Front ^[6]	Config = t1
3	RFC-0	Pad_length = 0
4	RFC-512	Pad_length = 512
5	RFC-1024	Pad_length = 1024
6	RFC-1460	Pad_length = 1460

本文构造的防御流量完全按照对应论文提供的代码生成，但由于RF Countermeasure源代码生成的流量不包含包长度信息，且未考虑包交互的时序性，本文对其代码进行了一些优化。在RF Countermeasure的源代码中，每个时隙(slot)内同一方向的数据包一次性发完，即每个时隙内恒有2个burst，这与实网情况不完全相符。当两个方向有延迟数据包时，可能会出现一侧包已达到发送上限，而另一侧包虽然没有达到上限，但受制于对侧的数据包发送终止，剩余部分在理论上仅能发送填充的虚拟数据包，而真实数据包应延迟到下一时隙。为了使混淆流量的结果尽量贴近于原始算法设计，本文对代码进行了有限的微调，将这部分额外的延迟数据包安排在下一时隙进行发送，其他情况与原方法一致。RF在RF Countermeasure防御流量上分类准确率为52.59%，但考虑包交互关系之后，准确率提高到74.2%，这说明官方代码忽略了数据包交互关系对分类的贡献，52.59%应是RF Countermeasure方法的理论最佳性能。

在复现RF Countermeasure时，使用不同的

虚拟包长度模拟不同的带宽消耗场景。Tor 协议最小数据单元 cell 的大小是 512 byte, 考虑 MTU 的限制, 本文选择 0 byte、512 byte、1024 byte、1460 byte 作为虚拟包长, 并依次将防御命名为 RFC-0、RFC-512、RFC-1024 和 RFC-1460。

3.2 参数选择

本节着重分析参数变化对方法性能的影响, 由于 ADF 主要解决 RF Countermeasure 防御方法带来的网站指纹识别困难问题, 因此模型的参数选择在 RF Countermeasure 防御场景下展开。主要包括包累积长度应用到单向/双向情况下性能的变化情况、超参数 T 、 s 和 N 的选择以及 ADF 在 RFC-0、RFC-512、RFC-1024 和 RFC-1460 等多种防御场景下性能的变化。

表 4 给出了在 RFC-0 防御下仅使用包数量、下行方向使用累积长度、上下行方向使用累积长度的分类准确率 (不含 burst 信息), 依次为 74.2%、86.8% 和 94.4%。结果显示累积包长度相较包数量对分类的贡献更大, 且双向使用效果最佳, 验证了 ADF 构造 HFM 时使用双向累积包长度的合理性。

表 4 累积包长度应用到单向/双向情况下准确率
Table 4 Accuracy of cumulative packet length applied to unidirectional/bidirectional case

参数	准确率
仅使用包数量	74.2%
下行用累积长度	86.8%
双向用累积长度	94.4%

RFC-0 防御下超参数 T 的取值对准确率的影响如图 6 所示。流量最大加载时间过短, 会使 HFM 损失后段流量的信息; 最大加载时间过长, 会使 HFM 变得稀疏, 分类精度下降的同时还会造成不必要的空间浪费。当最大加载时间取 $T = 80$ s 时, 分类准确率取得最大值, 因此本文取 $T = 80$ s。

RFC-0 防御下超参数 s 的取值对准确率的影响如图 7 所示。随着 s 的取值逐渐增大, 准确率整体呈现先上升后下降的趋势。时间窗口 s 过小, 会使防御方法对 HFM 元素的值扰动较明显; 时间窗口 s 过大, 会使 HFM 信息表示过于粗略, 分类精度下降。当 $s = 44$ ms 时, 分类准确率取

95.4%, 在保证抗干扰的同时兼顾了时间窗口内流特征的刻画粒度, 因此本文取 $s = 44$ ms。

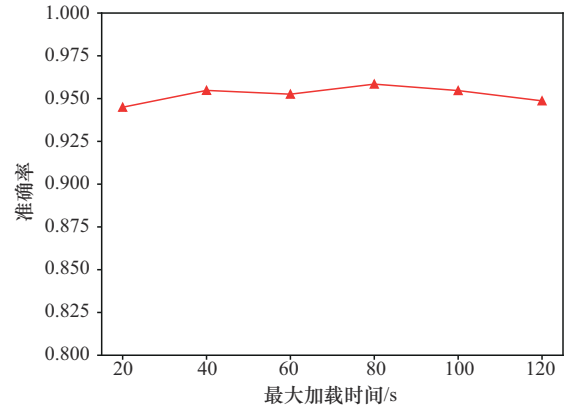


图 6 超参数 T 的取值对准确率的影响

Figure 6 Impact of hyperparameter T on accuracy

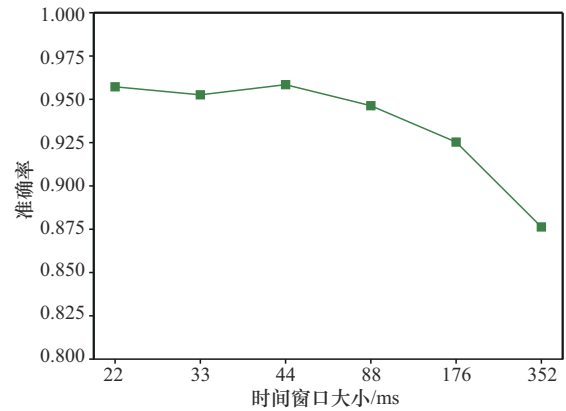


图 7 超参数 s 的取值对准确率的影响

Figure 7 Impact of hyperparameter s on accuracy

RFC-0 防御下超参数 N 的取值对准确率的影响如图 8 所示。2.1.2 节中提到, HFM 的第 3、4 行统计大小为 $N \times s$ 时间窗口内的 burst 数量, 其中 N 是一个超参数。由于 burst 属于粗粒度的特征, 时间窗口取值不宜过小, 故规定 N 取值下限为 2; 通过统计, 数据集中流的平均持续时间为 543 个 s (时间窗口), RF 模型的卷积核大小为 3×6 , N 取 90 即可让卷积核感受野覆盖整条流量, 故确定 N 取值上限为 90。结果显示, 对准确率的贡献主要来自 burst 数量, 与 N 无关, 本文取 $N = 9$ 。

ADF 在 RFC-0、RFC-512、RFC-1024 和 RFC-1460 上的准确率如图 9 所示。随着填充包长度增大, 对累积长度的扰动增强, 准确率有所下降, 但不低于 87.5%。

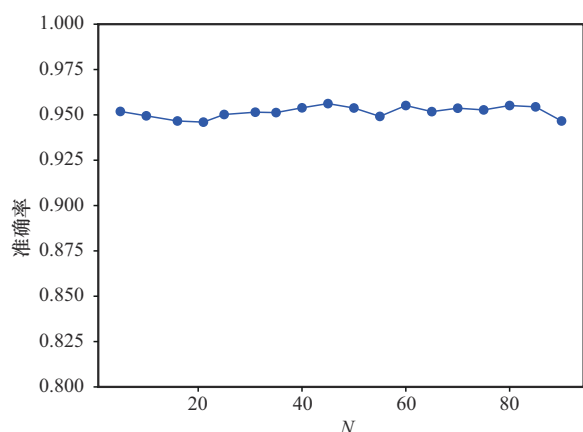
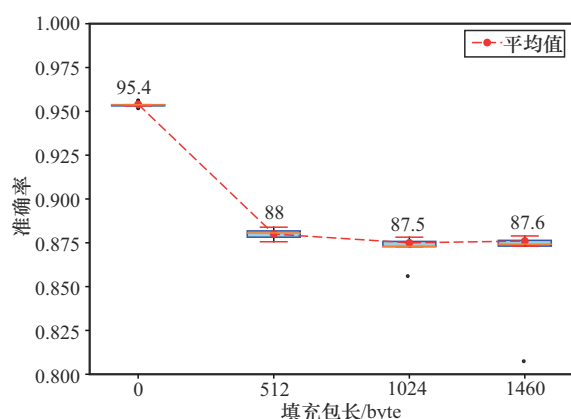
图8 超参数 N 的取值对准确率的影响Figure 8 Impact of hyperparameter N on accuracy of ADF

图9 填充包长变化对 ADF 准确率的影响

Figure 9 Impact of dummy packet length on accuracy of ADF

3.3 封闭世界场景下的实验

封闭世界采用准确率作为主要评价指标，即网站流量被正确划分为对应类别的准确率，计算方法如式(9)所示。

$$\text{Accuracy} = \frac{P_{\text{correct}}}{N} \quad (9)$$

其中， P_{correct} 为所有被正确分类的样本总数， N 为用于测试的样本总数。

为了进一步评价模型综合有效性，本文还采用了兼顾精确率 (precision) 和召回率 (recall) 的 $F1$ 值 ($F1\text{-score}$) 对方法的有效性进行验证：

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

其中， TP 表示真实类别为正例，预测类别为正例的样本数， FP 表示真实类别为负例，预测类别为正例的样本数， FN 表示真实类别为正例，预测类别为负例的样本数， TN 表示真实类别为负例，预测类别为负例的样本数。计算某一类的 TP 、 FP 、 FN 和 TN 时，将这一个类视为正类。

在多分类的问题中，为有效评估在多个类别中分类性能的整体性能，采用宏平均 (macro-averaging)，如式(13)所示。

$$\text{macro_F1} = \frac{1}{n} \sum_{i=1}^n F1_i \quad (13)$$

其中， n 表示网站类别数。

表5给出了本文所提方法 ADF 与 DF、Tik-Tok、RF 在 RF Countermeasure 防御条件下的准确率和宏平均值。数据结果表明，ADF 相较 DF、Tik-Tok 和 RF 在 RF Countermeasure 上的准确率分别提高了 35.7%、33.7% 和 21.2%，充分说明 ADF 方法的有效性。

表5 封闭世界 RF Countermeasure
防御场景下指标比较

Table 5 Metrics of RF Countermeasure defense in the closed-world scenario

方法	RFC-0*	
	准确率	macro_F1
DF ^[2]	59.7%	0.596
Tik-Tok ^[3]	61.7%	0.616
RF ^[7]	74.2%	0.773
ADF	95.4%	0.955

注：*这里仅使用 RFC-0 进行对比，因为对比方法皆未使用包长度特征，填充包长变化只会对 ADF 产生影响。

表6给出了 ADF 与 DF、Tik-Tok、RF 在无防御、WTF-PAD 和 Front 条件下的性能指标。结果表明，ADF 在 3 种防御条件下的准确率均高于 90%，与 RF 性能相当。

综上，在封闭世界场景下，ADF 相较以 RF 为代表的主流网站指纹识别方法，在 RF Countermeasure 防御上的分类准确率至少提升了 21.2%，同时在其他防御上仍保持良好的性能，验证了 ADF 的有效性和鲁棒性。

3.4 开放世界场景下的实验

由于监控集和非监控集的大小严重不平衡，

表6 封闭世界其他防御场景下指标比较

Table 6 Metrics of other defenses in the closed-world scenario

方法	无防御		WTF-PAD		Front	
	准确率	macro_F1	准确率	macro_F1	准确率	macro_F1
DF	98.3%	0.982	82.1%	0.822	82.6%	0.828
Tik-Tok	98.3%	0.983	86.4%	0.865	89.7%	0.897
RF	98.7%	0.988	93.6%	0.933	95.6%	0.955
ADF	98.1%	0.979	91.3%	0.913	94.9%	0.951

本文使用 P-R 曲线代替 TPR 和 FPR 作为开放世界的评价指标^[19-20], 计算方法如式(10)和式(11)所示。值得注意的是, 开放世界场景下变量的含义不同, TP 是监控集网站被正确分类的样本总数, TN 是非监控集网站被正确分类的样本总数, FP 是将非监控集中的网站误分为监控集网站的样本总数, FN 是将监控集中的网站误分为非监控集网站的样本总数。

网站指纹识别方可根据目标灵活调整识别策略。如果网站指纹识别方的目标是在高置信度下确定用户正在访问监控集网站, 则应追求高精确率; 如果网站指纹识别方的目标是广泛寻找可能访问监控集的用户, 则应追求高召回率。网站指纹识别方可通过调整阈值在两者之间进行权衡。

如图 10 所示, 在 RFC-0 数据集中, ADF 体现出明显的优势, 同等召回率下拥有最高的精确率, 同等精确率下拥有最高的召回率。

如图 11(a)所示, 在无防御数据集中, 4 种方法的性能都较为优秀。在高召回率下 ADF 和 RF 的精确率较高, ADF 的精确率略微低于 RF 的精

确率。如图 11(b)所示, 在 WTF-PAD 数据集中, 各方法的精确率从大到小排序依次为 RF、ADF、Tik-Tok、DF, 差距随着召回率增大而增大, 但总体精确率都在 90% 以上。如图 11(c)所示, 在 Front 数据集中, ADF 和 RF 的精确率优于 DF 和 Tik-Tok, 但 ADF 如果对精确率要求极高, 召回率会低于其他方法。

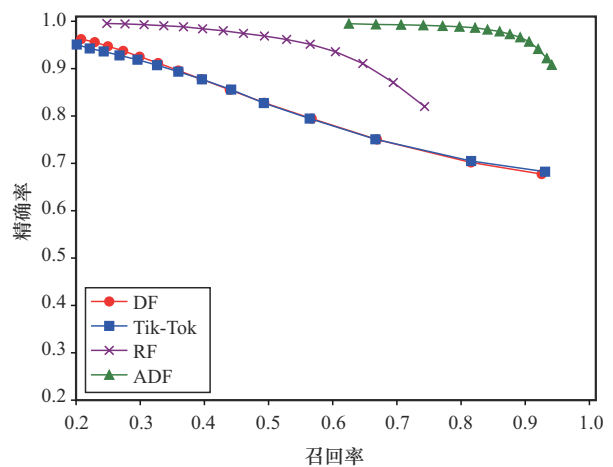


图 10 RFC-0 开放世界 P-R 曲线

Figure 10 Precision-recall curve of RFC-0 in the open-world

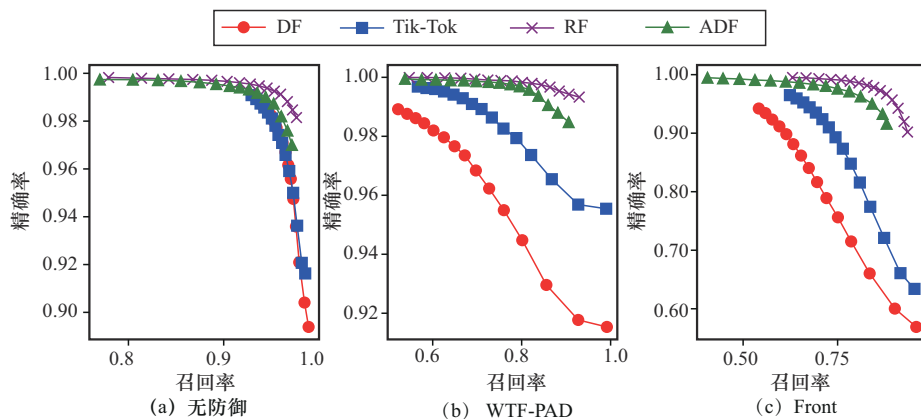


图 11 其他防御场景开放世界 P-R 曲线

Figure 11 Precision-recall curves of other defenses in the open-world

因此, ADF 在开放世界场景下对抗 RF Countermeasure 防御方法优于现有网站指纹识别方法, 对抗其他防御方法与 RF 性能相当, 证明了 ADF 的鲁棒性, 特别是在富信息域遭受干扰条件下能够取得较 RF 更优的性能。

4 结束语

为有效应对网站指纹防御方法广泛部署造成的流量特征偏移、网站类别混淆的现实挑战, 本文提出了抗多种防御策略的网站指纹识别方法 ADF。在使用包方向的基础上, 通过划分时间窗口、引入包累积长度和引入 burst 数量深度挖掘网站类别信息, 使 HFM 兼具局部特征和全局特征, 在对抗多样化防御扰动时具有更强的鲁棒性。通过大量实验分析发现, 在封闭世界和开放世界场景中, ADF 均能够在多样化防御扰动中有效分类, 在 RF Countermeasure 中其分类性能优于现有最好方法。

下一步, 笔者将针对未知网站流量分类、网站流量特征漂移等现实问题, 进一步研究增量更新、自主学习的网站指纹识别方法, 以期解决网站流量数据动态变化条件下模型高效更新的问题。

参考文献:

- [1] 邹鸿程, 苏金树, 魏子令, 等. 网站指纹识别与防御研究综述[J]. 计算机学报, 2022, 45(10): 2243-2278.
- [2] ZOU H C, SU J S, WEI Z L, et al. A review of the research of website fingerprinting identification and defense[J]. Chinese Journal of Computers, 2022, 45(10): 2243-2278.
- [3] SIRINAM P, IMANI M, JUAREZ M, et al. Deep fingerprinting: undermining website fingerprinting defenses with deep learning[C]// Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2018: 1928-1943.
- [4] RAHMAN M S, SIRINAM P, MATHEWS N, et al. Tik-tok: the utility of packet timing in website fingerprinting attacks[J]. Privacy Enhancing Technologies, 2020, 2020(3): 5-24.
- [5] BHAT S, LU D, KWON A, et al. Var-CNN: a data-efficient website fingerprinting attack based on deep learning[J]. Privacy Enhancing Technologies, 2019, 2019(4): 292-310.
- [6] JUAREZ M, IMANI M, PERRY M, et al. Toward an efficient website fingerprinting defense[C]// Proceedings of European Symposium on Research in Computer Security. Cham: Springer, 2016: 27-46.
- [7] GONG J, WANG T. Zero-delay lightweight defenses against website fingerprinting[C]// Proceedings of 29th USENIX Security Symposium (USENIX Security), 2020: 717-734.
- [8] SHEN M, JI K, GAO Z, et al. Subverting website fingerprinting defenses with robust traffic representation[C]// Proceedings of 32nd USENIX Security Symposium (USENIX Security 23). 2023: 607-624.
- [9] XIAO X, ZHOU X, YANG Z Y, et al. A comprehensive analysis of website fingerprinting defenses on Tor[J]. Computers & Security, 2024, 136: 103577.
- [10] WANG T, CAI X, NITHYANAND R, et al. Effective attacks and provable defenses for website fingerprinting[C]// Proceedings of 23rd USENIX Security Symposium (USENIX Security 14). 2014: 143-157.
- [11] PANCHENKO A, LANZE F B, ZINNEN A, et al. Website fingerprinting at Internet scale[C]// Proceedings of 2016 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2016: 1-16.
- [12] HAYES J, DANEZIS G. K-fingerprinting: a robust scalable website fingerprinting technique[EB]. arXiv Preprint arXiv: 1509.00789, 2015.
- [13] 赵文博, 马紫彤, 杨哲. 基于超图神经网络的恶意流量分类模型[J]. 网络与信息安全学报, 2023, 9(5): 166-177.
- [14] ZHAO W B, MA Z T, YANG Z. Model of the malicious traffic classification based on hypergraph neural network[J]. Chinese Journal of Network and Information Security, 2023, 9(5): 166-177.
- [15] 谢绒娜, 马铸鸿, 李宗俞, 等. 基于卷积神经网络的加密流量分类方法[J]. 网络与信息安全学报, 2022, 8(6): 84-91.
- [16] XIE R N, MA Z H, LI Z Y, et al. Encrypted traffic classification method based on convolutional neural network[J]. Chinese Journal of Network and Information Security, 2022, 8(6): 84-91.
- [17] SIRINAM P, MATHEWS N, RAHMAN M S, et al. Triplet fingerprinting: more practical and portable website fingerprinting with N-shot learning[C]// Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2019: 1131-1148.
- [18] OH S E, MATHEWS N, RAHMAN M S, et al. Gandalf: gan for data-limited fingerprinting[C]// Proceedings of Privacy Enhancing Technologies, 2021: 305-322.
- [19] WANG T, GOLDBERG I. {Walkie-Talkie}: An efficient defense against passive website fingerprinting attacks[C]// Proceedings of 26th USENIX Security Symposium (USENIX Security 17). 2017: 1375-1390.
- [20] LI S, GUO H J, HOPPER N. Measuring information leakage in website fingerprinting attacks and defenses[C]// Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2018: 1977-1992.

- [18] SHMATIKOV V, WANG M H. Timing analysis in low-latency mix networks: attacks and defenses[C]// Proceedings of European Symposium on Research in Computer Security (ESORICS), 2006: 18-33.
- [19] JUAREZ M, AFROZ S, ACAR G, et al. A critical evaluation of website fingerprinting attacks[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2014: 263-274.
- [20] PANCHENKO A, NIESSEN L, ZINNEN A, et al. Website fingerprinting in onion routing based anonymization networks[C]//Proceedings of the 10th annual ACM Workshop on Privacy in the Electronic Society. New York: ACM, 2011: 103-114.

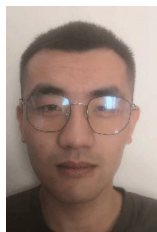
[作者简介]



张静茜 (2002-), 女, 云南昆明人, 信息工程大学硕士生, 主要研究方向为流量分析、匿名流量识别。



李腾耀 (1991-), 男, 河北辛集人, 博士, 信息工程大学讲师, 主要研究方向为匿名流量识别、去匿名化。



涂宇宽 (2000-), 男, 河南信阳人, 信息工程大学博士生, 主要研究方向为流量分析。



罗向阳 (1978-), 男, 湖北荆门人, 博士, 信息工程大学教授、博士生导师, 主要研究方向为网络与信息安全。