洋葱路由器网站指纹攻击与防御研究综述

杨宏宇*①② 宋成瑜② 王 朋② 赵永康② 胡 泽① 成 翔③④ 张 良⑤

①(中国民航大学安全科学与工程学院 天津 300300)

②(中国民航大学计算机科学与技术学院 天津 300300)

③(扬州大学信息工程学院 扬州 225127)

④(中国民航大学民航信息安全评估中心 天津 300300)

⑤(亚利桑那大学信息学院 图森 85721)

摘 要：以洋葱路由器(Tor)为代表的匿名网络是目前使用最广泛的加密通信网络之一，违法分子利用加密网络以掩盖其违法行为，给网络监管和网络安全带来极大的挑战。网站指纹攻击技术的出现使得对加密流量的分析成为可能，监管者利用数据包方向等信息对Tor流量进行解密，推断用户正在访问的网页。该文对Tor网站指纹攻击与防御方法进行了调研和分析。首先，对Tor网站指纹攻击的相关技术进行总结与比较，重点分析基于传统机器学习和深度学习的Tor网站指纹攻击；其次，对目前多种防御方法进行全面调研和分析；针对现有Tor网站指纹攻击方法存在的局限性进行分析和总结，展望未来发展方向和前景。

关键词：Tor匿名网络；网站指纹攻击；流量分析；隐私保护；网络监管

中图分类号：TN915.08; TP393 **文献标识码**：A **文章编号**：1009-5896(2024)09-3474-16

DOI: 10.11999/JEIT240091

Website Fingerprinting Attacks and Defenses on Tor: A Survey

YANG Hongyu①② SONG Chengyu② WANG Peng② ZHAO Yongkang②

HU Ze① CHENG Xiang③④ ZHANG Liang⑤

①(*School of Safety Science and Engineering, Civil Aviation University of China, Tianjin* 300300, *China*)

②(*School of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*)

③(*School of Information Engineering, Yangzhou University, Yangzhou* 225127, *China*)

④(*Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China, Tianjin* 300300, *China*)

⑤(*School of Information, The University of Arizona, Tucson* 85721, *USA*)

Abstract: The anonymity network represented by The onion router(Tor) is one of the most widely used encrypted communication networks, criminals utilize encrypted networks to conceal their illegal activities, posing significant challenges to network regulation and cybersecurity. The emergence of website fingerprinting attack has made the analysis of encrypted traffic possible, enabling supervisors to identify Tor traffic and infer the web pages being visited by users by utilizing features such as packet direction and so on. In this paper, a wide survey and analysis of website fingerprinting attack and defense methods on Tor are conducted. Firstly, relevant techniques of website fingerprinting attacks on Tor are summarized and compared. The emphasis is placed on website fingerprinting attacks based on traditional machine learning and deep learning technologies. Secondly, a comprehensive survey and analysis of various existing defense methods are conducted. The limitations in the field of website fingerprinting attack methods on Tor are analyzed and summarized, and the

收稿日期：2024-02-22；改回日期：2024-04-29；网络出版：2024-05-17

*通信作者： 杨宏宇 yhyxlx@hotmail.com

基金项目：国家自然科学基金(62201576, U1833107)，江苏省基础研究计划自然科学基金青年基金(BK20230558)，国家自然科学基金配套基金(3122023PT10)

Foundation Items: The National Natural Science Foundation of China (62201576, U1833107), Jiangsu Provincial Basic Research Program Natural Science Foundation——Youth Fund (BK20230558), The Supporting Fund of the National Natural Science Foundation of China (3122023PT10)

future development directions and prospects are looked forward to.

Key words: The onion router (Tor) anonymity network; Website fingerprinting attacks; Traffic analysis; Privacy protection; Network regulation

1　引言

大规模网络监控和互联网审查制度使得用户的行为隐私受到了严重侵犯，人们开始使用超文本传输安全协议(HyperText Transfer Protocol Secure, HTTPS)、安全外壳协议(Open Secure SHell, OpenSSH)和虚拟专用网络(Virtual Private Network, VPN)等加密网络寻求隐私保护。洋葱路由器(The onion router, Tor)[1]作为目前最流行的匿名网络之一，通过多个中继节点传输用户的通信数据，隐藏路由信息和通信内容对用户隐私进行保护。与此同时，Tor的匿名性质也导致它容易被不法分子用来进行违法活动，如发表不正当言论、部署僵尸网络及暗网交易等[2]。因此，有必要采取措施对Tor网络行为进行监督和审查。

网站指纹攻击属于加密流量分析技术，攻击者通过监测传输数据包的大小、方向、时间戳等信息来推断用户正在访问的网站。网站指纹攻击技术可以有效地对网络隐私行为进行监管，维护网络环境安全。但从用户隐私和安全的角度来看，网站指纹攻击的实际应用也可能被视为对个人隐私的侵犯。

目前，在网站指纹攻击的综述中，已有的工作包括：Sun等人[3]仅从单标签和多标签网站指纹攻击研究进行综述，未涉及防御研究且对局限性的讨论不足；Zou等人[4]对网站指纹攻击与防御方法进行了全面的综述，但并未将研究重点放在Tor网站指纹攻击领域。针对防御方法的划分存在冗余，多个类别间差异不够明显，对于实验设计和数据集的选择方面介绍较少；Shen等人[5]针对基于机器学习方法的加密流量分析方法进行了详细综述，但未深入探讨具体的实施细节与方法比较。

本文专注于Tor场景下的网站指纹攻击与防御，与现有研究相比，本文重点从实验方法、特征选择、实验数据集等方面详细介绍Tor网站指纹攻击，按照不同模型在封闭世界和开放世界中的表现进行分析和比较，揭示各种方法在不同应用环境下的适应性及其效能。针对Tor网站指纹防御，本研究将收集到的成果主要分为随机化防御、正则化防御、对抗性防御等，从防御效果、优缺点、可部署性等方面对现有防御方法进行归纳和总结。此外，本文对Tor网站指纹识别领域存在的局限性进行深入分析与总结，针对现有方法的局限性提出未来研究方向。

2　网站指纹攻击概述

网站指纹攻击的起源可以追溯到Wagner和Schneier[6]在1996年的一项研究，研究发现加密的HTTP请求所传递的数据包总量可能会暴露用户所访问的网站。随后研究人员在HTTPS[7]，OpenSSH[8]和VPN[9]等弱加密系统上开展了网站指纹的研究，取得了显著的成就。直到2009年，Herrmann等人[10]首次在Tor网络上成功实施了网站指纹攻击，尽管该方法准确率仅为3%。但随后Panchenko等人[11]使用支持向量机技术改进了该方法，将准确率提高至50%以上，开启了针对Tor网络的研究热潮。

2.1　Tor

Tor提供多种措施来保证用户的隐私和安全。首先，Tor以固定单元大小进行传输数据，通过数据填充技术，使攻击者难以根据数据包大小推测传输文件。其次，Tor客户端在随机选择的多个中继电路上传输信息，不同电路的客户端在延迟、拥塞和带宽容量等方面拥有不同的性能。这些因素会导致同一站点的数据包序列不同，增加网站指纹攻击难度。此外，Tor进行多项后台活动，如电路构建、速度测试、使用数据包进行流量控制等，对流量信息产生噪声干扰。最后，Tor使用流水线技术和数据包的随机发送来增加网络流量的不可预测性[12]。

2.2　威胁模型

在网站指纹攻击的威胁模型中，本地攻击者通过观察网络流量模式来推断Tor用户访问的页面。攻击者往往位于入口节点，可以是用户本地网络的窃听者、本地系统管理员、互联网服务提供商、或入口节点的运营商等。攻击者只能够观察和记录通过网络的流量痕迹，并没有能力丢弃、延迟或修改流量中的真实数据包，也不能对其进行解密[13]。Tor网站指纹攻击威胁模型如图1所示。

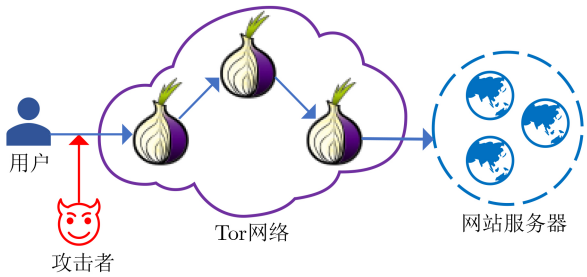

图 1 Tor网站指纹攻击威胁模型

2.3 防御模型

在威胁模型中，攻击者通常位于Tor入口节点，为了防止攻击者推测用户流量，防御模型必须操纵用户初始发送的流量。由于流量是双向的，成功部署防御方案需要客户端和服务器端共同参与，通常将防御方案设置在用户终端，或作为浏览器插件集成到浏览器内部[4]。因此，监管者最初监听到的流量已经经过混淆处理，难以辨识。基本防御策略包括添加虚拟数据包或延迟真实数据包的发送，但这种防御措施往往具有较高的数据开销(需要许多虚拟数据包)或对用户流量造成严重延迟，因此并未被Tor项目组所采用[14]。

3 Tor网站指纹攻击

研究人员针对Tor网站指纹攻击技术提出了大量方法，根据目标网站的类型，将其划分为单标签和多标签网站指纹攻击。本节重点讨论单标签网站指纹攻击，根据攻击方法的不同，进一步将其分为基于传统机器学习的方法和基于深度学习的方法。单标签Tor网站指纹攻击方法分类如图2所示。

3.1 基于传统机器学习的方法

早期Tor网站指纹攻击研究常采用自收集的数据集，不同方法之间收集的数据规模各不相同，环境也存在差异，导致不同方法之间的性能难以进行直接比较。因此本文从数据规模和准确率等多个方面综合评估各分类器的性能，基于传统机器学习的Tor网站指纹攻击模型的总结比较如表1所示。

3.1.1 基于朴素贝叶斯的方法

朴素贝叶斯(Naive Bayes, NB)是广泛使用的监督学习方法，该方法基于特征之间的独立性假设，认为每个特征在给定类别下相互独立，利用贝叶斯定理计算样本属于每个类别的概率，从而进行分类。

Liberatore和Levine[8]将Tor网站指纹攻击看作流量分析技术，首先将朴素贝叶斯引入到加密流量分析领域。基于Liberatore和Levine[8]的工作，Herrmann等人[10]引入多项式朴素贝叶斯算法对网站指纹进行分类，在安全外壳协议(OpenSSH)数据集上达到96.65%准确率。

Dyer等人[20]提出基于朴素贝叶斯的n元语法模型(Variable N-Gram, VNG)，使用总时间、总带宽和Burst等粗粒度特征来进行网站指纹攻击。研究发现粗粒度特征更能够帮助分类器判别经过数据包填充后的流量。

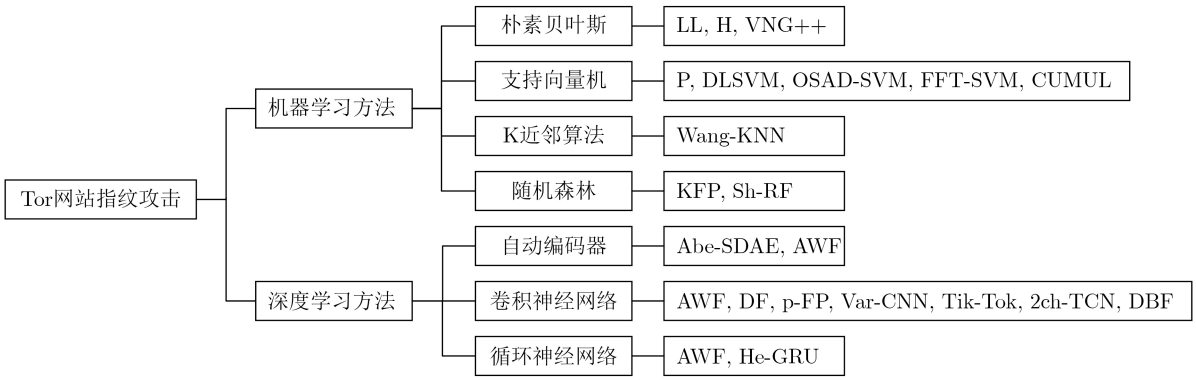

图 2 单标签Tor网站指纹攻击方法框架

表 1 基于传统机器学习的Tor网站指纹攻击模型比较

分类器	模型	数据单元	特征	封闭世界		开放世界		
				数据集规模	准确率(%)	数据集规模	TPR(%)	FPR(%)
H[10]	MNB	TCP/IP	带方向包长及计数	775×4	2.96			
P[11]	SVM	TCP/IP	带方向包长及计数等	775×20	54.61	4 000×1	73.00	0.05
DLSVM[15]	SVM	TCP/IP	数据包总数	100×40	83.70			
OSAD-SVM[12]	SVM	Cell	Cell方向	100×40	91.00	860×1	96.90	0.20
FFT-SVM[16]	SVM	TCP/IP	数据包大小、方向	100×40	>95.00			
CUMUL[17]	SVM	Cell	Cell大小、方向、顺序	100×90	91.38	9 000×1	96.64	9.61
Wang-KNN[18]	KNN	Cell	数据包长度等3 736个特征	100×90	91.00	5 000×1	85.00	0.60
KFP[19]	RF, KNN	TCP/IP	数据包统计特征	55×100+30×80	91.00	30×80+16 000×1	81.00	0.02

注：分类器命名规则采用已发表文献中使用的名称，若无统一名称，则根据作者名称的首部两个字符结合分类器类型自拟。

3.1.2 基于支持向量机的方法

支持向量机(Support Vector Machine, SVM)通过在多维特征空间中寻找最大间隔分离超平面来实现分类，分离超平面仅由少量支持向量决定。

Panchenko等人[11]首次将SVM应用于网站指纹攻击，通过引入流量的Burst特征，将Tor网站的分类准确率从3%提高至55%，在开放世界模型中实现了73%的召回率，首次在开放世界场景下取得成功。

研究人员进一步研究不同距离衡量方法对SVM分类效果的改进，Cai等人[15]使用Damerau-Levenshtein距离作为相似度度量。Wang等人[12]在Cai等人[15]的基础上，使用最佳字符串对齐距离(Optimal String Alignment Distance, OSAD)比较两个流量实例之间的差距。针对基于莱文斯坦距离设计的支持向量机(Damerau-Levenshtein Support Vector Machine, DLSVM)方法和OSAD-SVM方法中对于相似距离的计算均存在较高时间复杂度等问题，Jahani等人[16]引入快速傅里叶变换(Fast Fourier Transform, FFT)的方法计算流量实例的相似距离，在时间开销上比之前工作降低约400倍。

在特征选择方面，Panchenko等人[17]采用先前工作[11,18]中所提出的4个基本特征，包括传入、传出数据包的数量，以及传入、传出数据包大小的总和，并且采样100个额外的附加特征，使用带有径向基函数(Radial Basis Function, RBF)内核的开源SVM工具包[21](Library for Support Vector Machines, LibSVM)构建累积(CUMULative, CUMUL)模型。CUMUL选择的104个特征能够达到91.38%的精度，而Wang等人[18]所选择的3 736个特征仅达到90.84%的精度，进一步证实影响攻击效果的因素在于所选择特征的质量而不是数量。

3.1.3 基于K近邻算法的方法

K近邻算法(K-Nearest Neighbor, KNN)根据待分类样本周围最近的K个邻居所属类别来决定该样本的类别。该方法核心思想在于，单个样本如果被多数同类别的样本包围，则很可能属于这个类别。

Wang等人[18]基于KNN的机器学习方法，手动选择了大量特征，包括数据包排序、传入和传出数据包数量以及突发数量等，将特征组合起来形成距离度量来比较不同网站之间的相似性。用于评估的特征集包含近4 000个特征，实现了91%的准确率。

3.1.4 基于随机森林的方法

随机森林(Random Forest, RF)是由决策树集合组成的分类算法，结合了多个决策树模型，通过对预测结果的综合来提高模型的准确性和泛化能力。

Hayes等人[19]基于随机森林的集成学习方法，提出K指纹模型(K-FingerPrinting, KFP)。该模型利用随机森林方法评估特征的重要性，选择150个关键特征作为模型的输入，采用KNN方法对其进行分类。研究表明，相比于数据包排序或到达间隔时间等复杂特征，数据包数量等简单特征对网页身份的判别更为有效。与Hayes等人[19]的研究类似，Shen等人[22]也将随机森林应用于网站指纹的特征选择问题，提出一系列特征选择及优化的系统化方法。

3.2 基于深度学习的方法

基于深度学习的网站指纹攻击已经成为研究热点问题，本文收集了目前基于深度学习Tor网站指纹攻击研究工作，从分类器的模型、数据集以及表现性能等方面进行比较分析，结果如表2所示。

3.2.1 基于自动编码器的方法

Abe等人[23]首次将深度学习技术引入到Tor网站指纹攻击领域，提出基于堆叠去噪自动编码器(Stacked Denoising AutoEncoder, SDAE)的网站指纹攻击方法。实验仅使用Tor信元的方向作为输入向量，在封闭世界中取得88%的准确率。

Rimmer等人[24]系统性地将堆叠去噪自动编码器(SDAE)、卷积神经网络(Convolutional Neural Network, CNN)和长短期记忆网络(Long Short-Term Memory, LSTM)应用于Tor网站指纹攻击领域，提出自动化网站指纹攻击模型(Automated Website Fingerprinting, AWF)。Rimmer等人还收集了一个超过3×10^6个网络痕迹组成的数据集(记为Rimmer18)，这也是迄今为止最大规模，使用量最多的网站指纹数据集。在封闭世界场景下，随着网站数量的增多，SDAE分类器可以最大程度地保持检测性能。该实验首次证实深度学习自动提取的特征优于当前手工特征，展示出极大的发展空间。

3.2.2 基于卷积神经网络的方法

与Rimmer等人的工作非常相似，Oh等人[25]在预测指纹模型(predicting Fingerprintability, p-FP)中探索了多种深度学习架构的表现，其中包括多层感知机(MultiLayer Perceptron, MLP)、CNN、自动编码器(AutoEncoder, AE)。该研究首次指出当最先进的网站指纹攻击模型应用AE提取的特征时，其性能优于使用手动选取的特征集，实现更好的分类效果。

Sirinam等人[13]进一步改进了Rimmer等人[24]的模型，通过设计更为复杂的CNN来进行特征提取和分类，提出深度网站指纹攻击(Deep Fingerprinting, DF)模型。DF模型增加了卷积层数量并施加过拟合防护措施，针对Tor网站的攻击效果显

表 2 基于深度学习的Tor网站指纹攻击模型比较(%)

分类器	模型	数据单元	特征	数据集	封闭世界性能	开放世界性能	
					准确率	TPR	FPR
Abe-SDAE[23]	SDAE	Cell	Cell方向	Wang16	88.00	86.00	2.00
AWF_SDAE[24]	SDAE	Cell	Cell方向	Rimmer18	94.25	71.30	3.40
AWF_CNN[24]	CNN	Cell	Cell方向	Rimmer18	91.79	70.94	3.82
AWF_LSTM[24]	LSTM	Cell	Cell方向	Rimmer18	88.04	53.39	3.67
p-FP_MLP[25]	MLP	Cell	Cell序列、Burst	WTT		90.00	1.00
p-FP_CNN[25]	CNN	Cell	Cell序列、Burst	WTT		94.00	2.00
DF[13]	CNN	Cell	Cell方向	Sirinam18	98.30	95.70	0.70
Var-CNN[26]	CNN	Cell	Cell方向、时间戳	Rimmer18	98.80	98.01	0.36
Tik-Tok[27]	CNN	TCP/IP、TLS、Cell	Burst、原始时间序列	Sirinam18	98.40	94.00	
DBF[28]	CNN	Cell	Cell方向、Burst	Rimmer18	98.31	98.44	1.70
DBF[28]	CNN	Cell	Cell方向、Burst	Sirinam18	98.77	99.00	6.76
DBF[28]	CNN	Cell	Cell方向、Burst	Hayes16	70.60	68.42	0.57
2ch-TCN[29]	CNN	Cell	Cell方向、时间戳	Wang14	93.73		
WF-Transformer[30]	Transformer	Cell	Cell方向、时序特征	Sirinam18	99.10	96.90	0.70
He-GRU[31]	ResNet, GRU	Cell	Cell方向	Rimmer18	99.85	84.25	

著提升，准确率高达98.3%。Bhat等人[26]同样对CNN进行改进优化，提出名为Var-CNN的网站指纹攻击模型。该方法采用ResNet18模型从原始Tor信元中自动提取特征，通过设计方向CNN和时间CNN并将特征进行融合后取得了超越DF的效果。

先前基于深度学习的网站指纹攻击研究中，大多数工作并未充分利用数据包的时序信息。Rahman等人[27]提出一组基于Burst的时间特征，采用Sirinam等人[13]提出的DF模型来验证这些特征的有效性。根据数据包方向的不同将传出数据包表示为+1，传入数据包表示为–1，再令每个数据包的时间戳乘以其方向形成完整特征输入到DF分类器。在没有任何防御的Tor流量场景下，Tik-Tok模型达到98.4%的分类准确率，比仅使用方向信息的模型效果更优。Ma等人[28]也通过深度挖掘Tor流量的Burst特征来提高准确率，该方法通过设计3个CNN网络对长度不同的Burst特征进行有效提取，提出基于深度分析Burst特征的网站指纹攻击模型(Deep Burst-analysis based website Fingerprinting attack, DBF)。他们的网络架构与DF类似，创新之处在于在第1层卷积运算时利用大小不同的卷积核对长度不同的Burst特征进行提取。为了充分挖掘数据包的时序信息，Wang等人[29]提出基于双通道时间卷积网络(2-channel Temporal Convolutional Networks, 2ch-TCN)的网站指纹攻击模型，同时从数据包方向和时间信息中提取特征。在Wang14数据集上的实验显示，引入时间特征后，模型的准确率从92.60%提升至93.73%，证实时间信息确实可以被攻击者利用，进而提高分类器性能。

基于CNN设计的攻击模型重点关注网站指纹的空间特征，并未关注到Tor流量的时间特征，在捕捉数据包序列之间的前后依赖关系方面存在限制。Zhou等人[30]针对该问题提出基于Transformer的网站指纹攻击模型(Website Fingerprinting Transformer, WF-Transformer)，该模型利用Transformer网络提取流量轨迹的时间特征，有效捕捉Tor流量序列之间的长期依赖关系，在输入长度上和收敛速度上均优于基于CNN设计的模型。

3.2.3 基于循环神经网络的方法

相较于AE和CNN在网站指纹攻击领域的出色表现，循环神经网络通常表现不佳且具有较高的时间复杂度。然而循环神经网络却能够有效捕捉网站指纹的时序特征，通常与CNN配合使用。Rimmer等人[24]的方法中采用LSTM技术对数据集进行评估，与方法中的其他两种分类器相比，精度最低且执行速度最慢。He等人[31]提出基于残差网络(Residual neural Network, ResNet)和门控循环单元(Gated Recurrent Unit, GRU)的深度学习模型，采用双层GRU网络提取网站指纹的时间特征，使用ResNet-50网络提取网站指纹的空间特征，将两种特征融合后进行分类，达到99%以上的分类准确率。

3.3 多标签网站指纹攻击

在现实世界场景中，用户通常会连续打开多个

网页，多选项卡浏览引起的流量重叠问题往往将导致网站指纹模型无法正确分类。针对多标签网站指纹攻击的问题，Xu等人[32]采用平衡级联分割算法(BalanceCascade)来判别页面之间的分割点，并使用随机森林算法对分割后的页面块进行分类。Yin等人[33]进一步改进了Xu等人[32]的工作，通过判别第1页与其后续页面之间的分割点，提取第1个网页的初始干净数据块，仅根据这些数据块来对网站进行分类。该方法有效地绕过了分析或区分不同页面以及处理混合网络流量的复杂性问题。Gu等人[34]利用BalanceCascade算法分离重叠部分流量，基于ResNet和多头自注意力机制构建攻击模型。尽管现有多标签网站指纹攻击方法通常整个浏览会话划分为多个干净的流量块，每个块仅包含单一网站的流量，根据每个块推断访问的网站。然而客户端打开的选项卡数量往往是先验未知的，Deng等人[35]将多标签网站指纹攻击视作多标签分类问题，使用多分类器框架进行分类，从根本上放松对会话中所打开的选项卡数量的先验知识。

3.4 常用数据集

研究人员针对Tor网站指纹攻击收集了多个数据集，供学者进行测试和评估。早期的网站指纹攻击研究多采用机器学习技术，通常采取自收集数据的方法，数据规模较小且分类精度不佳。随着深度学习技术的出现和应用，研究人员开始在大规模数据集上评估自己的方法。本文收集了近年来单标签Tor网站指纹攻击领域常用的数据集，并与采用该数据集的相关研究方法进行关联，汇总整理至表3。

4 Tor网站指纹防御

为防御网站指纹攻击者，可在用户端或通过Tor网络部署防御策略，通过修改客户端与网络代理的通信方式来增强隐私保护。研究人员针对Tor网站指纹防御提出了大量方法，本文将搜集到的最新防御方法归类为随机化防御、正则化防御、对抗性防御以及其他防御方法，Tor网站指纹防御方法的分类如图3所示。

4.1 随机化防御

为防御网站指纹攻击，研究人员提出了多种随机化防御方法，通过在网页流量中随机填充虚拟数据包来进行流量混淆。然而此类方法具有较大的带宽开销，研究者们往往需要在防御开销和防御性能之间取得平衡。网站指纹防御方法的比较往往通过分类器的准确率下降程度、带宽开销和延迟来衡量。本文对现有的随机化防御方法进行了总结和比较，如表4所示。

随机化防御通常涉及向真实数据中填充虚拟数据包。Liberatore等人[8]首先提出数据包填充的方法来防御网站指纹攻击，他们提出四种虚拟字节填充策略，包括线性填充、指数填充、鼠象填充和最大传输单元(Maximum Transmission Unit, MTU)填充。

Liberatore等人[8]的方法在面对新型攻击模型时表现不佳，促使研究者们开发更为先进的防御方

表 3　单标签Tor网站指纹攻击常用数据集

数据集名称	数据集规模		研究方法关联
	封闭世界	开放世界	
Cai12[15]	100×40		DLSVM, FFT-SVM
Wang13[12]	100×40	860×1	OSAD-SVM, FFT-SVM
Wang14[18]	100×90	9 000×1	Wang-KNN, TF, AdaWFPA, 2ch-TCN
Wang16[36]	100×40	5 000×1	Abe-SDAE, p-FP
ALEXA100[17]	100×40	860×1	CUMUL, Sh-RF
Hayes16[19]	55×100+30×80	100 000×1	KFP, DBF
Rimmer18[24]	900×2 500	400 000×1	AWF, Var-CNN, TF, 2ch-TCN, DBF, He-GRU, snWF
Sirinam18[13]	95×1 000	40 716×1	DF, Tik-Tok, TF, DBF, WF-Transformer

图 3 Tor网站指纹防御方法

表 4 随机化防御方法比较

防御名称	防御效果		优点	缺点
	分类器模型	准确率变化(%)		
WTF-PAD[37]	P	55.00→15.33	轻量级防御，无通信延迟	混淆时间特征的能力有限
	DLSVM	83.70→23.00		
FRONT[14]	Wang-KNN	83.18→41.22	轻量级防御，专注于混淆跟踪前端，随机化虚拟数据包的数量和分布，无通信延迟	混淆时间特征的能力有限
	CUMUL	64.22→11.97		
	KFP	94.38→71.19		
	DF	91.12→34.88		
Camouflage[11]	P	55.00→3.00	无需对匿名网络进行任何修改，易于实施	对部分页面仍然无法提供有效保护
RanDePad[38]	KFP	89.98→54.15	具备低且可控的带宽开销	未设计延迟控制方案
	CUMUL	90.77→50.39		
	DF	94.57→62.40		

法。自适应网站指纹防御(Website Traffic Fingerprinting Protection with Adaptive Defense, WTF-PAD)[37]采用改进的自适应填充方法来防御攻击模型。为使填充更加真实，该方法根据防御服务器接收到的消息进行反馈来发送填充消息，模拟HTTP的请求响应流程，进一步扭曲流量的Burst特征。网络流量的前沿混淆(Front Randomized Obfuscation of Network Traffic, FRONT)[14]则通过高度随机的方式添加虚拟数据包，确保同一网页的不同实例在总长度、数据包顺序和数据包方向上看起来彼此不同，以此来进行网络流量混淆。FRONT[14]和WTF-PAD[37]专注于零延迟填充技术，使用随机的虚拟数据包填充网站流量，不对真实流量数据包进行任何更改。然而，零延迟填充技术由于无延迟地发送真实的数据包，将会泄漏流量数据包的时间信息。针对该问题，Hong等人[38]提出自适应随机延迟和填充(adaptive Random Delaying and Padding, RanDePad)防御方法，采用随机延迟技术来破坏网站流量的时间分布特征。该方法通过带宽评估算法对流量带宽进行评估，动态调整带宽填充方案以隐藏流量空间特征，确保低且可控的带宽开销。

Panchenko等人[11]提出背景噪声的方法来防御先前提出的P分类器。该方法在每次加载页面时，通过随机加载多个网页来混淆数据流量。所提方法可以通过网络浏览器插件来实现，易于实施且无需对匿名服务Tor或Java匿名代理(Java Anon Proxy, JAP)进行任何修改。

4.2 正则化防御

正则化防御方法通过定义所有网页流量必须遵循的固定规则和模式来限制客户端发送和接收数据包的方式，严格限制攻击者可用特征空间[39]。本节将各种正则化防御方法进行比较，并总结归纳至表5。

尽管Tor匿名服务对数据包大小进行填充，但攻击者仍然可以通过分析流量的粗略特征(如总时间和总带宽)来准确地推断用户访问的网站。针对目前防御存在的缺陷，Dyer等人[20]提出缓冲区固定长度混淆(Buffered Fixed Length Obfuscation, BuFLO)防御方法，该方法将报文按照固定时间间隔、固定长度、固定速率发送，消除可能用于流量分析的侧信道信息。但在实际应用时存在发送效率低下，延迟开销大等问题。针对BuFLO防御存在的问题，Cai等人[40]提出了改进的拥塞敏感(Congestion Sensitive-BuFLO, CS-BuFLO)防御方法，该方法增加了拥塞敏感和自适应速率等功能，能够有效地隐藏网站总大小和最后一个对象的大小等宏观特征，同时能够根据网络的拥塞情况动态调整传输速率，减少带宽的延迟和浪费。Cai等人[41]进一步将数据包大小设置为750字节而不是MTU，将传出流量固定为比传入流量更高的数据包间隔，以减少带宽和时间开销，提出Tamaraw防御方法。

固定速率填充的防御方法需要确保数据包具有相同结束时间，因此在时间开销上面临较高的延迟，在实际应用过程中可能会受到用户诟病，研究人员进一步提出流量变形的方法，通过空间开销来缓解时间延迟。Wang等人[18]通过构建最短公共超序列(Supersequence)来保护网络通信的隐私，通过计算匿名集中数据包序列的最短公共超序列，代替原始的数据包序列进行通信，达到隐藏真实通信内容的目的。Gong等人[14]提出的粘合轨迹(GlUE)防御方法通过在多个页面访问之间添加虚拟数据包，使客户端呈现出一种连续不断地访问新页面的假象，掩盖实际行为特征。

Wang等人[42]限制浏览器以半双工模式进行通信，确保数据包的时序、方向和顺序完全相同，以

表 5　正则化防御方法比较

防御名称	防御效果		优点	缺点
	分类器模型	准确率变化(%)		
BuFLO[20]	H	2.96→0.80	严格限制攻击者可利用的特征空间	效率极低，粗粒度特征仍然会泄露网站有关信息
	P	54.61→27.30		
CS-BuFLO[40]	P	54.61→23.40	具有拥塞敏感和自适应速率等功能，减少带宽对BuFLO进行改进以隐藏最重要的流量特征，开销可调节	延迟较高
	DLSVM	83.70→＜30.00		
Tamaraw[41]				重量级防御，延迟较高
Supersequence[18]	Wang-KNN	91.00→6.80	可得到最优防御的输出包序列	需要先验知识，选择最短公共超序列问题面临较大计算复杂度
GLUE[14]	Wang-KNN	83.18→＜5.00	攻击者需要对页面进行分割，难度较大。用户可根据选择定制开销	延迟较高
	CUMUL	64.22→＜5.00		
	KFP	94.38→＜5.00		
	DF	91.12→＜5.00		
Walkie-Talkie[42]	P	81.00→44.00	带宽可调节，灵活且开销极低	前提需要知道用户将要访问网页的一些信息。需要修改浏览器加载网页的方式，部署困难
	DLSVM	94.00→19.00		
	OSAD-SVM	97.00→25.00		
	Wang-KNN	95.00→28.00		
	CUMUL	64.00→20.00		
	KFP	86.00→41.00		
RegulaTor[43]	Tik-Tok	97.00→25.40	带宽开销最小，且不需要额外的基础设施或其他跟踪知识	需要延迟数据传输
	DF	98.40→19.60		
	CUMUL	97.20→16.30		

较低的额外开销向对手泄露更少的信息，提出对讲机(Walkie-Talkie)防御方法。该方法将敏感网页的数据包序列转换为非敏感网页，确保两种数据包序列完全一致，混淆攻击者，然而该方法却显著增加了页面加载时间。Liang等人[44]的工作详细分析了页面加载时间增加的原因，提出了名为尾部时间(Tail Time, TT)的防御方法，该方法通过限制待处理请求以及定时发送所有被阻止的请求来减少页面加载时间，取得了与Walkie-Talkie相同的防御性能。

现有防御措施要么会因延迟增加而影响用户体验，要么因带宽增加而给Tor网络带来负担，要么需要创建和维护额外的基础设施。Holland等人[43]针对以上问题提出正则化洋葱路由(RegulaTor)防御方法。该方法通过正则化下载流量中常见的激增数据包的大小和形状，使用衰减的目标速率来控制数据发送速度，以此来避免泄露有关激增的流量和长度的信息。

4.3　对抗性防御

深度学习已被证明容易受到对抗样本的影响[45]，通过对原始样本进行微小的扰动，从而误导分类器做出错误的判断。本节针对研究人员在对抗性防御方面所做的工作进行分析和比较，归纳总结至表6，其中部分防御效果通过扰动成功率来衡量。

对抗样本常被用来混淆深度学习模型，但如果攻击者利用现有防御方法生成对抗痕迹，并用它们训练更为强大的分类器，则可能造成更大的威胁。为了抵抗经过对抗性训练增强的攻击模型，Qiao等人[51]开发一种名为Acup3的黑盒防御方法。该方法通过模仿多个网站的流量，使不同网站的流量痕迹看起来更为相似，增加了分类的难度。其次，Acup3无需访问流量跟踪即可生成与跟踪无关的扰动，使其适合实际部署。最后，该方法通过扰动变异来多样化不同用户访问同一网站的流量轨迹，有效地限制了经过对抗性训练的模型。Rahman等人[46]提出一种生成对抗性痕迹的防御方法知更鸟(Mockingbird)，该方法通过增加搜索过程的随机性以及减少对目标深度学习模型的依赖，生成难以通过对抗性训练找到的路径。

Hou等人[47]提出基于生成对抗网络的网站指纹防御方法(Website Fingerprinting attack-Generative Adversarial Network, WF-GAN)，该方法通过训练将源网站特征映射到对抗样本，使得对抗样本特征逼近目标网站。Gong等人[48]同样基于生成

表 6 对抗性防御方法比较

防御名称	防御效果		优点	缺点
	分类器模型	准确率变化(%)		
Mockingbird[46]	DF	97.00→38.00	限制对抗性训练的有效性	需要提前了解完整的流量突发序列，实时性较差，部署性较差
	Var-CNN	98.00→30.00		
	CUMUL	93.00→20.00		
	KFP	85.00→26.00		
	Wang-KNN	86.00→12.00		
WF-GAN[47]	DF	90.00(扰动成功率)	同时具备无针对性和有针对性的防御能力	需要提前了解完整的流量突发序列，部署性较差
Surakav[48]	KFP	73.62→0.01	采用多种发送模式，实时灵活调整，开销较小	产生一定程度的拥塞，影响数据包调度
	CUMUL	74.23→2.74		
	DF	96.24→8.14		
	Tik-Tok	96.68→6.28		
Blind[49]	DF	92.00→1.00	能够有效防御盲目对抗性攻击	模型训练时间较长，无法扩展到更大的模型
	Var-CNN	93.00→1.40		
Minipatch[50]	AWF	83.60(扰动成功率)	带宽消耗更低，防御性能更好	不能防御使用时间特征的网指纹攻击模型
	DF	60.90(扰动成功率)		
	Var-CNN	70.50(扰动成功率)		

对抗网络提出名为Surakav的防御方法，该方法设计了一个生成对抗网络来模拟网页的真实发送模式。

研究人员还通过对抗扰动技术对真实数据包进行扰动来达到防御目的。Nasr等人[49]提出盲目性对抗扰动技术来击败现有的基于DNN的流量分析系统。扰动包括更改数据包的时间和大小，以及插入虚拟网络数据包等。Li等人[50]基于对抗扰动的思想，通过向网络流量中注入极少的虚拟数据包来破坏基于DNN的网站指纹攻击模型，提出Minipatch防御方法。该方法使用对抗性补丁来扰乱网站跟踪，诱导分类器进行错误分类。Gu等人[52]利用梯度加权类激活映射(Gradient-weighted Class Activation Mapping, Grad-CAM)算法来查明在分类过程中发挥重要作用的关键部分，通过分段和施加约束来生成较强的对抗性补丁，使之能够在实时交通场景中注入和删除预先训练的补丁。

从现有防御方法来看，大多数方法对协议栈和网络条件做出简化假设，有些条件甚至是实际环境或应用时所不具备的，因此无法真实反映防御的性能。Gong等人[53]提出在Tor网络上创建可插拔传输的通用平台WFDefProxy，实现目前所有已知的网络层防御方法。该平台通过简化防御实施过程，允许防御开发人员在更现实的环境中对其方法进行评估，简化了基准测试流程。需要注意的是，现有的对抗性防御方法主要针对特定的攻击模型，而忽略了其他潜在的攻击者，因此这些防御方法的实用性仍需进一步研究[54]。与此同时，未来网站指纹防御算法的可部署性仍然值得研究人员进一步探索。

4.4 其他防御

先前的防御方法大多基于流量混淆与正则化思想，还有少部分研究采取流量塑形和流量分割的方法进行防御。Wright等人[55]首次提出通过流量塑形(Traffic Morphing)的概念，通过向数据包添加填充或将数据包拆分为多个较小的单元，使客户端请求伪装成来自另一网页，进而阻止流量分析。但该方法在Tor上无效，因为Tor信元级的填充大小固定，数据包长度已经不会泄漏任何信息。

Cadena等人[56]提出流量分割(Traffic Sliver)的网络防御方法，通过在Tor网络中使用多路径通信和流量分割策略来增强用户的隐私保护。该方法将单个HTTP请求分割成不同的请求对象，分布到包含不同入口节点的多个Tor路径中，限制单个入口洋葱路由器的网络流量，扭曲网站指纹可重复利用流量模式。Liu等人[57]进一步提出名为SMART的轻量级网站指纹防御方法。该方法在Tor网络中引入多路径传输模型，将流量划分在多个Tor入口中继处，用户可以自主选择路径以优化匿名性和网络性能。此外，该方法引入冗余编码的概念，有效解决多径传输混沌到达和数据丢失问题，保证传输的可靠性。

5 现有方法的局限性

随着机器学习和深度学习技术的不断进步，网

站指纹攻击的精度值得到进一步提升，准确率开始接近理论极限值。研究人员开始关注到现有方法的局限性，包括识别假设、概念漂移等问题。此类问题在先前的工作中往往被研究人员所忽略，表面上较高的精度值实际是简化该场景下存在的问题所得到的结果，解决该问题可能会显著提高Tor网站指纹攻击在实际部署过程中的性能。

5.1 识别假设

现有的机器学习和深度学习技术在Tor网站指纹攻击上取得了出色表现，但多数研究工作中采取的识别假设极大地简化了问题，通过简化场景或高估对手的能力，给攻击者带来了不切实际的优势。Juarez等人[58]首次进行了批评封闭世界、顺序浏览、流量隔离和可复制性等假设的研究。本文将现有识别假设划分为3个部分，即用户行为假设、对手行为假设和网页假设。

5.1.1 用户行为假设

用户行为假设从攻击者角度出发，限定用户的浏览行为，假设主要包含封闭世界、顺序浏览和流量隔离假设。在封闭世界假设中，研究人员认为Tor用户仅访问攻击者预先收集的固定数量网页，并且所有这些网页都被用于训练分类器。封闭世界下的研究通常可以提供更高的准确率，但不适合现实场景，仅用作比较和分析不同分类器的性能。少数研究仅关注封闭世界场景下的性能，而大多数研究同时关注模型在封闭世界和开放世界的表现[59]。

在顺序浏览假设中，Tor用户以顺序且非连续的方式浏览网络，即用户1次只能访问1个网页，待当前网页加载完毕后才顺序浏览下一个网页[36]。大多数先前的工作假设Tor用户顺序浏览网站，且1次仅打开1个标签页，以便网站流量不会重叠。然而这并不能代表Tor客户端的真实行为，由于Tor的连接速度很慢，客户端很有可能会打开多个浏览器选项卡并同时访问站点。

在流量隔离假设中，Tor用户没有在网络上进行任何其他活动，如下载文件或更新操作系统等与互联网相关的后台活动[26]。在现实情况下，Tor用户很有可能在进行浏览页面的同时进行其他后台活动，后台活动产生的流量将显著干扰攻击过程。

5.1.2 对手行为假设

对手行为假设是对监管者攻击能力的假设，主要包含流量解析和可复制性假设。在流量解析假设中，监管者可以判别流量跟踪中页面加载的开始和结束，并且能够有效地分割重叠部分的流量。然而在实际应用中，即使页面加载之间存在较大时间间隔，对重叠流量的分类仍然是一项挑战。

在可复制性假设中，对手有能力在与受害者相同的条件下对其分类器进行训练。假设攻击者能够复制Tor用户的系统和网络环境，包括操作系统、网络连接以及Tor浏览器版本等。可复制性假设对于模拟用户浏览流量的模式至关重要，环境的差异可能显著影响浏览流量的模式，从而降低攻击模型的效果[26]。对手可以在相同配置下获取与用户相同的数据进行训练和测试，通过消除环境因素带来的差异，极大地提高了攻击性能。然而，在真实场景中，简单复制用户环境并非易事，特别是当攻击对象是多个受害者时[58]。

5.1.3 网页假设

网页假设主要是对用户所浏览的网页类型进行的假设，主要包含单页假设、静态网页、被动网页和禁用缓存假设。单页假设指研究人员通常假设监管者仅监视网站上的单个网页，该网页通常是主页。部分研究人员认为仅通过监控网站主页就足以确定用户是否浏览过网站，然而这种假设不够现实[12]，用户很可能同时浏览网站的主页和内部网页。

静态网页假设首次由Wang等人[12]提出，他们建议避免在内容频繁更新的网站上进行网站指纹攻击，例如每隔几分钟更新一次的新闻网站或视频内容网站。训练流量分类器的过程往往耗时过长，当网页内容更新时，攻击者必须使用新的浏览流量来训练分类器，代价十分昂贵。因此，攻击往往选择静态网页或定期更新内容的网站如政府机构的网站。

被动网页假设是指Tor用户在浏览网页时，与网页有关的所有主动内容都处于被禁用或非活动状态。Aminuddin等人[59]将主动内容定义为网页内容动态加载时的Flash, Java Applet或JavaScript流量痕迹。此外，与页面有关的广告、插件以及部分网页在特殊节日对主页进行的修改都应当被禁止，此类活动内容将导致同一个网页每次加载的流量模式都不同。与流量隔离假设不同，被动网页从网页自身出发，禁止了有可能产生流量痕迹的各种活动行为，而流量隔离是从攻击者角度出发，假设攻击者能够分离各种噪声。

研究人员假设使用Tor的用户将禁用浏览器缓存，以便每次浏览网页时，浏览器都会下载该网页的全部资源文件，避免访问任何本地缓存，确保多次访问也能够产生相同的流量模式。尽管Tor浏览器避免在浏览会话之间存储缓存，但在同一浏览会话期间仍然会利用缓存文件。大多数Tor上的网站指纹攻击都集中在用户仅访问主页1次的情况下，因此研究人员忽略了浏览器缓存所带来的影响。此外，不同的网站对缓存大小的配置不同，对缓存的

保存时间也存在显著差异，因此禁用缓存策略进一步简化了识别假设。

5.2 概念漂移

网站指纹攻击中的概念漂移是指网页内容随着时间的推移而不断变化的现象，导致流量轨迹的模式发生变化，影响模型分类的准确性。Juarez等人[58]首先揭示了网站内容更新对网站指纹攻击的影响，他们观察到随着时间的推移，网站指纹攻击的准确率急剧下降，在不到10天的时间内下降了50%，并且在90天后进一步下降到几乎为0。

Al-Naami等人[60]首次针对网站指纹攻击中的概念漂移问题，提出修复更新和动态更新方法。Attarian等人[61]系统调研了数据流挖掘算法在处理概念漂移问题时的能力，采用多种流算法包括自适应Hoeffding树[62]、自适应概念快速决策树[63](Concept adapting Very Fast Decision Tree, CVFDT)、OzaBag[64]和极快决策树[65](Extreme Fast Decision Tree, EFDT)在多分类任务上进行评估和比较。自适应Hoeffding树由于能够自动检测概念漂移并对自身参数进行更新从而拥有更高的分类准确率。此后，Attarian等人[66]进一步提出了自适应在线网站指纹攻击(Adaptive online Website FingerPrinting Attack, AdaWFPA)方法。该方法使用自适应Hoeffding树[62]算法对网站指纹攻击进行建模，并在每个训练实例到达时增量更新其模型，模型随着每个网站的最新版本不断更新，因此随着时间的推移将性能保持在最佳水平。然而，由于该过程涉及模型参数的不断更新，无法与现有深度学习模型的抗概念漂移性能进行直接比较。

Wang等人[67]提出一种基于神经网络集成技术的快照网站指纹攻击(snapshot Website Fingerprinting, snWF)，该方法采用较为经济性的快照集成策略，在测试阶段使用每个快照模型进行检测，最终结果由所有快照模型结果的平均值来决定。研究还发现，在概念漂移影响下，网站指纹攻击在开放世界环境中的性能下降更为严重。因此，在更现实的开放世界环境中研究概念漂移对攻击模型的影响仍然是一个值得解决的问题。

5.3 数据集局限性

为了保持较高的攻击准确率，攻击者通常需要收集有关网站的最新数据集，确保新的网站指纹能够被覆盖。然而现有的数据收集策略需要大量的时间和资源，并且随着时间的推移，网站指纹攻击的准确性会进一步下降。对于资源受限的攻击者而言，需要在限制监控集大小、使用陈旧的数据结果和每个站点使用较少的训练样本之间进行抉择[15]。

因此研究人员开始关注小样本学习在该领域的应用， Sirinam等人[68]首次将小样本学习(N-Shot Learning, NSL)引入到网站指纹攻击领域，提出三重指纹(Triplet Fingerprinting, TF)模型。该模型采用CNN进行特征提取，再使用KNN算法进行分类，仅使用每个网站的20个训练实例即可实现高达95%的分类准确率，极大减少了收集和训练大型网站指纹数据集的工作量。Chen等人[69]受迁移学习的启发，在攻击过程中引入了预先训练的特征提取器。然而该模型需要大量额外的预训练数据，当面对与预训练数据分布不同的目标数据时，提取器还有可能会失效。

Zhou等人[70]提出集群网站指纹攻击模型(Cluster Website Fingerprinting Attack, CWFA)，该模型基于集群假设，即属于同一集群的样本具有相同的类别。CWFA通过深度神经网络提取轨迹特征，并以标记样本的类别中心作为聚类中心，对未标记的目标轨迹的特征进行聚类。Chen等人[71]针对数据短缺问题，提出了一个称为发送和接收对(Sendand-Receive Pair, SRP)的概念来解析流量轨迹并设计基于SRP的累积特征。根据重新排列的SRP进一步重建和生成仿生痕迹，实验结果表明，所提仿生痕迹可以提高最先进的基于深度学习的Var-CNN的性能。

6 未来研究方向

随着Tor网站指纹攻击技术的不断发展，准确率得到不断优化，然而距离实际部署还存在一定的差距。目前该领域还存在假设性过强、概念漂移问题严重、数据不足等局限性。本文针对现有方法的局限性及挑战进行总结并展望，未来解决好这些问题有助于提高网站指纹攻击模型在现实环境中的实用性和可靠性。

6.1 弱化识别假设

弱化识别假设是网站指纹攻击领域未来研究的一个重要方向。通过减少对用户、对手和网页行为的识别假设，并考虑更多实际场景中的因素，可以提高网站指纹攻击模型的实用性和鲁棒性，满足实际应用的需求。

针对用户行为假设，未来研究可以考虑设计更为灵活的用户行为模型，即允许用户同时打开多个标签页和进行后台活动等，模拟真实用户的网络行为。针对对手行为假设，未来可以探索更多、更复杂的网站指纹攻击方法，同时模拟用户环境进行训练。针对网页假设，在未来研究过程中可以进一步扩展到更多类型的网页，包括内部网页，频繁更新

的网页，噪声较多的网页及缓存网页等，增加模型对各种网页类型的适应性和分析能力。

现有大多数网站指纹攻击研究集中于受控环境或实验室实验，有意或无意地采用部分假设以简化其攻击过程，缺乏对其所采用假设的透明度。部分隐含的假设条件，甚至是在实际环境中或应用过程中所不具备的。未来可以深入研究各类假设对网站指纹攻击模型的影响，并对其工作中所应用的假设保持透明，将有助于提高整体网站指纹攻击研究在现实世界中的可行性。

6.2 缓解概念漂移问题

在网站指纹攻击模型中，分类器采用预先收集的数据集进行训练。随着时间的推移，网站流量轨迹的变化并未体现在训练后的分类器中，分类准确率显著降低。现有大部分研究工作仅停留在采用深度学习提取更为通用、稳定的网站指纹特征来抗衡概念漂移问题，在攻击效果上达到瓶颈。

在未来研究过程中，可引入模型或数据更新技术实现对的分类器的可持续更新。针对模型更新，采用增量学习或在线学习等方式将最新的网站流量轨迹纳入训练过程，避免重新训练整体模型，提高模型更新效率和实时性要求。针对数据更新，可持续监测和收集网站指纹的最新数据，引入滑动窗口对数据进行划分，定期丢弃旧批次数据。其次，可引入集成学习等方法，通过构建多个不同的网络模型来降低单个神经网络的偏差和方差，增强模型的鲁棒性，提升模型在概念漂移数据上的表现。

6.3 增强数据多样性

现有的数据集在开放世界场景下，仅包含有限数量的主页，缺乏对其内部网页的收集，这使得网站指纹模型可能会失效。实际上，用户浏览不受任何限制，即使在同一网站内，用户也极有可能访问多个不同的内部页面。在未来的研究过程中，可以在各种非监控集和内部网页上收集数据，调查网站指纹攻击模型的有效性。

此外，现有工作缺乏对网页缓存策略的系统化研究，往往采用禁用缓存策略来简化假设。近期，Karunanayake等人[72]发现，在Tor浏览器缓存的影响下，重新加载网页产生的流量与首次访问网页时的网站指纹显著不同，这一发现挑战了现有模型的攻击效果。在未来的数据收集过程中，应当包括具备不同缓存策略的网页资源，深入分析缓存策略对网站指纹攻击效果的影响。

6.4 流量分割

现有大部分多标签网站指纹攻击模型需要先验了解客户端所打开的标签数量，并在给定固定数量的选项卡情况下进行训练。当标签数量动态变化且先验未知时，此类模型难以进行有效分类[71]。在未来研究工作过程中，应当进一步放宽对标签数量的限定，设计更为合理和更具有部署性的多标签网站指纹攻击模型。

现有方法能够有效处理两个或多个浏览网页之间的正时间间隔和零时间间隔问题，针对负时间间隔以及完全重叠流量轨迹段的分离，仍然缺乏有效的分割算法。未来针对连续浏览或重叠浏览的网页进行流量分割值得研究人员进一步探索。

6.5 防御场景下的网站指纹攻击

为了降低网站指纹攻击效果，研究人员提出多种防御方法，主要分为随机化防御、正则化防御、对抗性防御等。最近的网站指纹攻击模型开始利用各种流量表示(如数据包方向、时间戳等)和深度学习技术来破坏现有的防御，但各种攻击仍无法针对不同的防御方法实现较高的攻击准确率。因此，在未来研究工作中，如何提升模型的普适性及鲁棒性从而克服现有防御方案值得研究人员进一步研究。最近Shen等人[73]的工作专注于防御场景下的网站指纹攻击，通过设计一种强大的流量表示，充分捕获从Tor跟踪中泄漏的关键信息特征而不受现有防御方法的干扰，在多个防御数据集上超越现有模型。

目前各类防御方法大多因为通信负担和延迟而受到用户诟病，各类防御方法大多没有被Tor项目组所采用。随着时间的发展，Tor项目组也将随之推出各类防御方案来保护用户隐私，未来针对防御场景下的网站指纹攻击仍然值得研究。

7 结论

随着网络加密技术的不断发展，互联网安全与隐私面临巨大挑战。Tor作为最具代表性的匿名网络之一，一方面为用户提供隐私保护服务，另一方面也成逐渐为犯罪分子进行网络违法活动的平台。本文收集并分析了Tor网站指纹攻击领域最新的研究成果，将现有的攻击方法主要划分为基于传统机器学习的方法与基于深度学习的方法。此外，本文对Tor网站指纹的防御方法进行了深入调查和分类，从多个维度对各防御方法进行分析和比较。最后，本文对当前Tor网站指纹攻击和防御方法的局限性进行归纳总结，展望未来的研究方向。

参 考 文 献

[1] DINGLEDINE R, MATHEWSON N, and SYVERSON P F. Tor: The second-generation onion router[C]. The 13th USENIX Security Symposium, San Diego, USA, 2004: 303–320.

[2] KARUNANAYAKE I, AHMED N, MALANEY R, *et al*. De-anonymisation attacks on Tor: A survey[J]. *IEEE Communications Surveys & Tutorials*, 2021, 23(4): 2324–2350. doi: 10.1109/COMST.2021.3093615.

[3] 孙学良, 黄安欣, 罗夏朴, 等. 针对Tor的网页指纹识别研究综述[J]. 计算机研究与发展, 2021, 58(8): 1773–1788. doi: 10.7544/issn1000-1239.2021.20200498.
SUN Xueliang, HUANG Anxin, LUO Xiapu, *et al*. Webpage fingerprinting identification on Tor: A survey[J]. *Journal of Computer Research and Development*, 2021, 58(8): 1773–1788. doi: 10.7544/issn1000-1239.2021.20200498.

[4] 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. doi: 10.11897/SP.J.1016.2022.02243.
ZOU Hongcheng, SU Jinshu, WEI Ziling, *et al*. A review of the research of website fingerprinting identification and defense[J]. *Chinese Journal of Computers*, 2022, 45(10): 2243–2278. doi: 10.11897/SP.J.1016.2022.02243.

[5] SHEN Meng, YE Ke, LIU Xingtong, *et al*. Machine learning-powered encrypted network traffic analysis: A comprehensive survey[J]. *IEEE Communications Surveys & Tutorials*, 2023, 25(1): 791–824. doi: 10.1109/COMST.2022.3208196.

[6] WAGNER D and SCHNEIER B. Analysis of the SSL 3.0 protocol[C]. The 2nd USENIX Workshop on Electronic Commerce, Oakland, USA, 1996: 4. doi: 10.5555/1267167.1267171.

[7] HINTZ A. Fingerprinting websites using traffic analysis[C]. The 2nd International Workshop on Privacy Enhancing Technologies, San Francisco, USA, 2003: 171–178. doi: 10.1007/3-540-36467-6_13.

[8] LIBERATORE M and LEVINE B N. Inferring the source of encrypted HTTP connections[C]. The 13th ACM Conference on Computer and Communications Security, Alexandria, USA, 2006: 255–263. doi: 10.1145/1180405.1180437.

[9] BISSIAS G D, LIBERATORE M, JENSEN D, *et al*. Privacy vulnerabilities in encrypted HTTP streams[C]. The 5th International Workshop on Privacy Enhancing Technologies, Cavtat, Croatia, 2006: 1–11. doi: 10.1007/11767831_1.

[10] HERRMANN D, WENDOLSKY R, and FEDERRATH H. Website fingerprinting: Attacking popular privacy enhancing technologies with the multinomial naïve-Bayes classifier[C]. The ACM Workshop on Cloud Computing Security, Chicago, USA, 2009: 31–42. doi: 10.1145/1655008.1655013.

[11] PANCHENKO A, NIESSEN L, ZINNEN A, *et al*. Website fingerprinting in onion routing based anonymization networks[C]. The 10th Annual ACM Workshop on Privacy in the Electronic Society, Chicago, USA, 2011: 103–114. doi: 10.1145/2046556.2046570.

[12] WANG Tao and GOLDBERG I. Improved website fingerprinting on tor[C]. The 12th ACM Workshop on Privacy in the Electronic Society, Berlin, Germany, 2013: 201–212. doi: 10.1145/2517840.2517851.

[13] SIRINAM P, IMANI M, JUAREZ M, *et al*. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning[C]. The 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, Canada, 2018: 1928–1943. doi: 10.1145/3243734.3243768.

[14] GONG Jiajun and WANG Tao. Zero-delay lightweight defenses against website fingerprinting[C]. The 29th USENIX Conference on Security Symposium, Berkley, USA, 2020: 41. doi: 10.5555/3489212.3489253.

[15] CAI Xiang, ZHANG Xincheng, JOSHI B, *et al*. Touching from a distance: Website fingerprinting attacks and defenses[C]. The ACM Conference on Computer and Communications Security, Raleigh, USA, 2012: 605–616. doi: 10.1145/2382196.2382260.

[16] JAHANI H and JALILI S. A novel passive website fingerprinting attack on tor using fast Fourier transform[J]. *Computer Communications*, 2016, 96: 43–51. doi: 10.1016/j.comcom.2016.05.019.

[17] PANCHENKO A, LANZE F, PENNEKAMP J, *et al*. Website fingerprinting at internet scale[C]. The 23rd Annual Network and Distributed System Security Symposium, San Diego, USA, 2016: 1–15.

[18] WANG Tao, CAI Xiang, NITHYANAND R, *et al*. Effective attacks and provable defenses for website fingerprinting[C]. The 23rd USENIX Conference on Security Symposium, San Diego, USA, 2014: 143–157. doi: 10.5555/2671225.2671235.

[19] HAYES J and DANEZIS G. K-fingerprinting: A robust scalable website fingerprinting technique[C]. The 25th USENIX Conference on Security Symposium, Austin, USA, 2016: 1187–1203. doi: 10.5555/3241094.3241186.

[20] DYER K P, COULL S E, RISTENPART T, *et al*. Peek-a-boo, I still see you: Why efficient traffic analysis countermeasures fail[C]. The IEEE Symposium on Security and Privacy, San Francisco, USA, 2012: 332–346. doi: 10.1109/SP.2012.28.

[21] CHANG C C and LIN C J. LIBSVM: A library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): 27. doi: 10.1145/1961189.1961199.

[22] SHEN Meng, LIU Yiting, ZHU Liehuang, *et al*. Optimizing feature selection for efficient encrypted traffic classification: A systematic approach[J]. *IEEE Network*, 2020, 34(4):

20–27. doi: 10.1109/MNET.011.1900366.

[23] ABE K and GOTO S. Fingerprinting attack on Tor anonymity using deep learning[J]. *Proceedings of the Asia-Pacific Advanced Network, Hongkong, China*, 2016, 42: 15–20.

[24] RIMMER V, PREUVENEERS D, JUAREZ M, *et al*. Automated website fingerprinting through deep learning[C]. The 25th Network and Distributed System Security Symposium, San Diego, USA, 2018.

[25] OH S E, SUNKAM S, and HOPPER N. p^1-FP: Extraction, classification, and prediction of website fingerprints with deep learning[J]. *Proceedings on Privacy Enhancing Technologies*, 2019, 2019(3): 191–209. doi: 10.2478/popets-2019-0043.

[26] BHAT S, LU D, KWON A, *et al*. Var-CNN: A data-efficient website fingerprinting attack based on deep learning[J]. *Proceedings on Privacy Enhancing Technologies*, 2019, 2019(4): 292–310. doi: 10.2478/popets-2019-0070.

[27] RAHMAN M S, SIRINAM P, MATHEWS N, *et al*. Tik-Tok: The utility of packet timing in website fingerprinting attacks[J]. *Proceedings on Privacy Enhancing Technologies*, 2020, 2020(3): 5–24. doi: 10.2478/popets-2020-0043.

[28] 马陈城, 杜学绘, 曹利峰, 等. 基于深度神经网络burst特征分析的网站指纹攻击方法[J]. 计算机研究与发展, 2020, 57(4): 746–766. doi: 10.7544/issn1000-1239.2020.20190860.
MA Chencheng, DU Xuehui, CAO Lifeng, *et al*. Burst-analysis website fingerprinting attack based on deep neural network[J]. *Journal of Computer Research and Development*, 2020, 57(4): 746–766. doi: 10.7544/issn1000-1239.2020.20190860.

[29] WANG Meiqi, LI Yanzeng, WANG Xuebin, *et al*. 2ch-TCN: A website fingerprinting attack over tor using 2-channel temporal convolutional networks[C]. The IEEE Symposium on Computers and Communications, Rennes, France, 2020: 1–7. doi: 10.1109/ISCC50000.2020.9219717.

[30] ZHOU Qiang, WANG Liangmin, ZHU Huijuan, *et al*. WF-transformer: Learning temporal features for accurate anonymous traffic identification by using transformer networks[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 30–43. doi: 10.1109/TIFS.2023.3318966.

[31] HE Xiaomin, WANG Jin, HE Yueying, *et al*. A deep learning approach for website fingerprinting attack[C]. The 4th International Conference on Computer and Communications, Chengdu, China, 2018: 1419–1423. doi: 10.1109/CompComm.2018.8780755.

[32] XU Yixiao, WANG Tao, LI Qi, *et al*. A multi-tab website fingerprinting attack[C]. The 34th Annual Computer Security Applications Conference, San Juan, USA, 2018: 327–341. doi: 10.1145/3274694.3274697.

[33] YIN Qilei, LIU Zhuotao, LI Qi, *et al*. An automated multi-tab website fingerprinting attack[J]. *IEEE Transactions on Dependable and Secure Computing*, 2022, 19(6): 3656–3670. doi: 10.1109/TDSC.2021.3104869.

[34] GU Xiaodan, YANG Ming, SONG Bingchen, *et al*. A practical multi-tab website fingerprinting attack[J]. *Journal of Information Security and Applications*, 2023, 79: 103627. doi: 10.1016/j.jisa.2023.103627.

[35] DENG Xinhao, YIN Qilei, LIU Zhuotao, *et al*. Robust multi-tab website fingerprinting attacks in the wild[C]. 2023 IEEE Symposium on Security and Privacy, San Francisco, USA, 2023: 1005–1022. doi: 10.1109/SP46215.2023.10179464.

[36] WANG Tao and GOLDBERG I. On realistically attacking tor with website fingerprinting[J]. *Proceedings on Privacy Enhancing Technologies*, 2016, 2016(4): 21–36. doi: 10.1515/popets-2016-0027.

[37] JUAREZ M, IMANI M, PERRY M, *et al*. Toward an efficient website fingerprinting defense[C]. The 21st European Symposium on Research in Computer Security, Heraklion, Greece, 2016: 27–46. doi: 10.1007/978-3-319-45744-4_2.

[38] HONG Xueshu, MA Xingkong, LI Shaoyong, *et al*. A website fingerprint defense technology with low delay and controllable bandwidth[J]. *Computer Communications*, 2022, 193: 332–345. doi: 10.1016/j.comcom.2022.06.028.

[39] LIU Peidong, HE Longtao, and LI Zhoujun. A survey on deep learning for website fingerprinting attacks and defenses[J]. *IEEE Access*, 2023, 11: 26033–26047. doi: 10.1109/ACCESS.2023.3253559.

[40] CAI Xiang, NITHYANAND R, and JOHNSON R. CS-BuFLO: A congestion sensitive website fingerprinting defense[C]. The 13th Workshop on Privacy in the Electronic Society, Scottsdale, USA, 2014: 121–130. doi: 10.1145/2665943.2665949.

[41] CAI Xiang, NITHYANAND R, WANG Tao, *et al*. A systematic approach to developing and evaluating website fingerprinting defenses[C]. The 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, USA, 2014: 227–238. doi: 10.1145/2660267.2660362.

[42] WANG Tao and GOLDBERG I. Walkie-talkie: An efficient defense against passive website fingerprinting attacks[C]. The 26th USENIX Conference on Security Symposium, Vancouver, Canada, 2017: 1375–1390. doi: 10.5555/3241189.3241296.

[43] HOLLAND J K and HOPPER N. RegulaTor: A

straightforward website fingerprinting defense[J]. *Proceedings on Privacy Enhancing Technologies*, 2022, 2022(2): 344–362. doi: 10.2478/popets-2022-0049.

[44] LIANG Jingyuan, YU Chansu, SUH K, *et al*. Tail time defense against website fingerprinting attacks[J]. *IEEE Access*, 2022, 10: 18516–18525. doi: 10.1109/ACCESS.2022.3146236.

[45] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[C]. The 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[46] RAHMAN M S, IMANI M, MATHEWS N, *et al*. Mockingbird: Defending against deep-learning-based website fingerprinting attacks with adversarial traces[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 1594–1609. doi: 10.1109/TIFS.2020.3039691.

[47] HOU Chengshang, GOU Gaopeng, SHI Junzheng, *et al*. WF-GAN: Fighting back against website fingerprinting attack using adversarial learning[C]. The IEEE Symposium on Computers and Communications, Rennes, France, 2020: 1–7. doi: 10.1109/ISCC50000.2020.9219593.

[48] GONG Jiajun, ZHANG Wuqi, ZHANG C, *et al*. Surakav: Generating realistic traces for a strong website fingerprinting defense[C]. The IEEE Symposium on Security and Privacy, San Francisco, USA, 2022: 1558–1573. doi: 10.1109/SP46214.2022.9833722.

[49] NASR M, BAHRAMALI A, and HOUMANSADR A. Defeating DNN-based traffic analysis systems in real-time with blind adversarial perturbations[C]. The 30th USENIX Security Symposium, Vancouver, Canada, 2021: 2705–2722.

[50] LI Ding, ZHU Yuefei, CHEN Minghao, *et al*. Minipatch: Undermining DNN-based website fingerprinting with adversarial patches[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 2437–2451. doi: 10.1109/TIFS.2022.3186743.

[51] QIAO Litao, WU Bang, YIN Shuijun, *et al*. Resisting DNN-based website fingerprinting attacks enhanced by adversarial training[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 5375–5386. doi: 10.1109/TIFS.2023.3304528.

[52] GU Xiaodan, SONG Bingchen, LAN Wei, *et al*. An online website fingerprinting defense based on the non-targeted adversarial patch[J]. *Tsinghua Science and Technology*, 2023, 28(6): 1148–1159. doi: 10.26599/TST.2023.9010062.

[53] GONG Jiajun, ZHANG Wuqi, ZHANG C, *et al*. WFDefProxy: Real world implementation and evaluation of website fingerprinting defenses[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 1357–1371. doi: 10.1109/TIFS.2023.3327662.

[54] XIAO Xi, ZHOU Xiang, YANG Zhenyu, *et al*. A comprehensive analysis of website fingerprinting defenses on Tor[J]. *Computers & Security*, 2024, 136: 103577. doi: 10.1016/J.COSE.2023.103577.

[55] WRIGHT C V, COULL S E, and MONROSE F. Traffic morphing: An efficient defense against statistical traffic analysis[C]. The 16th Network and Distributed System Security Symposium, San Diego, USA, 2009.

[56] DE LA CADENA W, MITSEVA A, HILLER J, *et al*. TrafficSliver: Fighting website fingerprinting attacks with traffic splitting[C]. The 2020 ACM SIGSAC Conference on Computer and Communications Security, Orlando, USA, 2020: 1971–1985. doi: 10.1145/3372297.3423351.

[57] LIU Ling, HU Ning, SHAN Chun, *et al*. SMART: A lightweight and reliable multi-path transmission model against website fingerprinting attacks[J]. *Electronics*, 2023, 12(7): 1668. doi: 10.3390/electronics12071668.

[58] JUAREZ M, AFROZ S, ACAR G, *et al*. A critical evaluation of website fingerprinting attacks[C]. The 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, USA, 2014: 263–274. doi: 10.1145/2660267.2660368.

[59] AMINUDDIN M A I M, ZAABA Z F, SAMSUDIN A, *et al*. The rise of website fingerprinting on Tor: Analysis on techniques and assumptions[J]. *Journal of Network and Computer Applications*, 2023, 212: 103582. doi: 10.1016/j.jnca.2023.103582.

[60] Al-NAAMI K, CHANDRA S, MUSTAFA A, *et al*. Adaptive encrypted traffic fingerprinting with bi-directional dependence[C]. The 32nd Annual Conference on Computer Security Applications, Los Angeles, USA, 2016: 177–188. doi: 10.1145/2991079.2991123.

[61] ATTARIAN R and HASHEMI S. Investigating the streaming algorithms usage in website fingerprinting attack against Tor privacy enhancing technology[C]. The 16th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology, Mashhad, Iran, 2019: 33–38. doi: 10.1109/ISCISC48546.2019.8985162.

[62] PFAHRINGER B, HOLMES G, and KIRKBY R. New options for hoeffding trees[C]. The 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, 2007: 90–99. doi: 10.1007/978-3-540-76928-6_11.

[63] HULTEN G, SPENCER L, and DOMINGOS P. Mining time-changing data streams[C]. The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2001: 97–106. doi: 10.1145/502512.502529.

[64] OZA N C and RUSSELL S J. Online bagging and

boosting[C]. The Eighth International Workshop on Artificial Intelligence and Statistics, Key West, USA, 2001: 229–236.

[65] MANAPRAGADA C, WEBB G I, and SALEHI M. Extremely fast decision tree[C]. The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 2018: 1953–1962. doi: 10.1145/3219819.3220005.

[66] ATTARIAN R, ABDI L, and HASHEMI S. AdaWFPA: Adaptive online website fingerprinting attack for tor anonymous network: A stream-wise paradigm[J]. *Computer Communications*, 2019, 148: 74–85. doi: 10.1016/j.comcom.2019.09.008.

[67] WANG Yanbin, XU Haitao, GUO Zhenhao, *et al*. SnWF: Website fingerprinting attack by ensembling the snapshot of deep learning[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 1214–1226. doi: 10.1109/TIFS.2022.3158086.

[68] SIRINAM P, MATHEWS N, RAHMAN M S, *et al*. Triplet fingerprinting: More practical and portable website fingerprinting with N-shot learning[C]. The 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 2019: 1131–1148. doi: 10.1145/3319535.3354217.

[69] CHEN Mantun, WANG Yongjun, and ZHU Xiatian. Few-shot website fingerprinting attack with meta-bias learning[J]. *Pattern Recognition*, 2022, 130: 108739. doi: 10.1016/j.patcog.2022.108739.

[70] ZHOU Qiang, WANG Liangmin, ZHU Huijuan, *et al*. Few-shot website fingerprinting attack with cluster adaptation[J]. *Computer Networks*, 2023, 229: 109780. doi: 10.1016/j.comnet.2023.109780.

[71] CHEN Yongxin, WANG Yongjun, and YANG Luming. SRP: A microscopic look at the composition mechanism of website fingerprinting[J]. *Applied Sciences*, 2022, 12(15): 7937. doi: 10.3390/app12157937.

[72] KARUNANAYAKE I, JIANG Jiaojiao, AHMED N, *et al*. Exploring uncharted waters of website fingerprinting[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 1840–1854. doi: 10.1109/TIFS.2023.3342607.

[73] SHEN Meng, JI Kexin, GAO Zhenbo, *et al*. Subverting website fingerprinting defenses with robust traffic representation[C]. The 32th USENIX Security Symposium, Anaheim, USA, 2023: 607–624.

杨宏宇：男，教授，博士生导师，研究方向为网络与系统安全、软件安全、网络安全态势感知.

宋成瑜：男，硕士生，研究方向为网络与系统安全、软件安全.

王　朋：男，助理教授，研究方向为大规模轨迹数据管理，城市计算，网络与系统安全.

赵永康：男，讲师，研究方向为信息安全，多媒体信息隐藏，秘密分享，视觉密码.

胡　泽：男，讲师，研究方向为人工智能、自然语言处理、网络信息安全.

成　翔：男，讲师，研究方向为网络与系统安全、网络安全态势感知、APT攻击检测.

张　良：男，研究员，研究方向为强化学习，基于深度学习的信号处理，网络与系统安全.

责任编辑：余　蓉