基于神经网络的恶意DNS流量检测方法

单康康[1]，袁书宏[1]，陈文智[1,2]，王志波[1,3]
（1.浙江大学信息技术中心，浙江 杭州 310027；
2.浙江大学计算机科学与技术学院，浙江 杭州 310027；
3.浙江大学网络空间安全学院，浙江 杭州 310027）

摘　要：针对目前机器学习检测恶意DNS流量提取流量特征方面的效率不高、检测准确率和检测速度较低等问题，提出了一种结合频域特征聚合分析和神经网络算法的恶意DNS流量检测方法FDS-DL。首先，通过离散傅里叶变换将DNS流量从时域空间转换到频域空间，在保留流量关键信息的同时大幅压缩数据规模；然后，利用卷积神经网络对处理后的频域序列数据进行分类。实验结果表明，与当前主流的几种检测方法相比，FDS-DL对恶意DNS流量的检测精度和F1_score性能最优。

关键词：频域；离散傅里叶变换；神经网络；卷积神经网络；恶意域名
中图分类号：TP309
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024232

Malicious DNS traffic detection based neural networks

SHAN Kangkang[1], YUAN Shuhong[1], CHEN Wenzhi[1,2], WANG Zhibo[1,3]
1. Information Technology Center, Zhejiang University, Hangzhou 310027, China
2. College of Computer Science, Zhejiang University, Hangzhou 310027, China
3. School of Cyber Science and Technology, Zhejiang University, Hangzhou 310027, China

Abstract: To solve the problems of low detection accuracy and speed caused by low efficiency in extracting traffic features using machine learning to detect malicious DNS traffic, a malicious DNS traffic detection method FDS-DL was proposed, which combines frequency domain feature aggregation analysis and neural networks algorithms. Firstly, DNS traffic was converted from time-domain space to frequency-domain space through discrete Fourier transform, which could significantly compress the data scale while retaining key log information. Then, convolutional neural network was used to classify the processed frequency domain sequence data. The experimental results show that compared with several mainstream detection methods, FDS-DL has a higher accuracy in identifying malicious DNS traffic and F1_score is optimal.

Keywords: frequency domain, DFT, neural network, convolutional neural network, malicious domain name

0　引言

随着互联网的快速发展，网络安全已经成为人们关注的重点话题之一。许多攻击者利用网络漏洞来进行各种恶意活动，例如钓鱼攻击、恶意软件传播、垃圾邮件发送等[1]。域名系统（DNS, domain name system）用于将主机名映射到IP地址，据统计，大约有一半以上的网络攻击都是利用DNS来进行的[2]。

收稿日期：2024–10–21
基金项目：未来互联网试验设施FITI项目试验节点建设资助项目（发改高技[2016]2533号）；中国高校产学研创新基金资助项目（No.2022HS046）
Foundation Items: Future Internet Experimental Facility FITI Project Experimental Node Construction (No. FGGJ[2016]2533), Industry-University-Research Innovation Fund for Chinese Universities (No.2022HS046)

机器学习的兴起极大地帮助了检测恶意域名。Grill等[3]、Gao等[4]使用特定的算法将未知域的网络行为与已知的网络恶意行为进行聚类，以确定这些域是恶意的还是良性的，尽管这种方法取得了良好的结果，但其鲁棒性较差，并且参数的调整对结果有很大影响。Schüppen等[5]、Casino等[6]、Alaei-yan等[7]结合域名字符特征和DNS流量特征对未知域名的签名进行评分以检测未知域名。上述方法在特定的数据集中都取得了良好的效果，但它们都是基于域名的低维特征。虽然这些方法中提取的特征具有多样性，但是特征的维度类同，并且上述方法只分析了部分恶意域名数据集，而没有研究未知域名之间的深层关系。当面对不断变化的恶意域名时，它们的有效性可能会大大降低。Zhang等[8]从DNS流量和日志提取特征并识别域名。尽管这些检测方法可以高精度地完成检测任务，但却为攻击者提供了一个机会，使他们可以在发起攻击时瞄准并更改此类检测方法所需的重要判断特征，例如域名存活了多久、响应IP是否包含多个域，从而大大降低了这些检测方法的泛化能力。为了检测和识别未知域名，在DNS流量级别进一步探索域名之间的相关性。Tran等[9]、Peng等[10]基于DNS流量构建域名之间的关系连接图模型，以检测未知域名是良性的还是恶意的，在DNS流量特征级别构建不同域名之间的关系，以检测和识别未知域名，虽然这些方法达到了很高的准确性，但是需要大量的计算，而且需要大量的资源来启动模型。此类方法中的大多数模型只能应用于静态场景，缺乏实时可用性。Yin等[11]使用业内权威域名列表作为算法的种子，同时对DNS流量中生成的“NXDOMAIN”响应进行抓取，对算法的输出值进行叠加，并将叠加值与设置的阈值进行比较，以检测主机是否连接到恶意的C&C服务器，该方法简单、成本低、易于部署，可用于计算能力有限的物联网设备，但是该方法仅使用“NXDOMAIN”响应DNS特征作为算法操作的符号，可能导致遗漏和误报，从而无法准确地对每个域名进行分类。此外，算法阈值的触发需要更多的恶意域，降低了该方法的及时性和可靠性。Sun等[12]认为目前大多数恶意域都是由域名生成算法（DGA, domain generation algorithm）生成的，因此研究在DGA域检测中引入可以学习样本间相关性并提取数据特征的机器学习模型。Woodbridge[13]等利用长短期记忆（LSTM）模型检测DGA域。Tran等[14]利用改进的LSTM对DGA域进行二元和多类分类。Vinayakumar等[15]比较多种机器学习模型组合对DGA域的二元和多类分类的有效性，机器学习模型在领域分类方面非常出色，但对于训练阶段尚未学习的DGA领域，它们的可移植性较差，并且无法准确预测未知领域的属性。

本文主要研究工作如下。

1) 提出了一种结合频域特征聚合分析和神经网络算法的恶意DNS流量检测方法，即FDS-DL，结合离散傅里叶变换（DFT）频域转换和卷积神经网络（CNN）对DNS流量进行分类。

2) 建立了良性域名选择列表，以减少手动选择参数的工作量，从而确保检测精度，同时避免手动设置参数。

3）利用离散傅里叶变换对DNS流量进行频域特征分析来提取序列信息，提高了后续的检测速度和精度。

4) 实验结果表明，FDS-DL方法具有很好的检测准确率和检测效率，并具有良好的性能。

1 FDS-DL检测框架模型

FDS-DL主要由5个阶段组成：时序处理、数据清洗、频域转换、数据归一化、神经网络训练与分类。FDS-DL方法模型结构如图1所示。

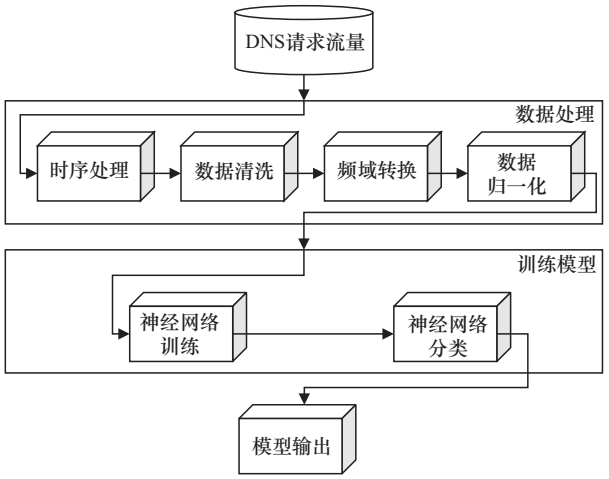

图1 FDS-DL方法模型结构

1) 时序处理阶段。将原始的DNS数据流特征序列编码转化为向量，以减少数据规模和后续处理的开销。对DNS流量进行时间排序可以帮助识别异常流量，因为网络中的正常流量通常具有一

定的时序关系。将网络流量按照时间排序的优点是便于发现某些时间相关的变化模式和规律，更好地挖掘潜在规律，更容易发现周期性的事件和攻击事件的发生，有助于识别异常情况。另外，还可以更准确地识别高峰时间等有意义的时间段。同时，按时间排序也有助于在时间序列分析中应用神经网络技术，包括时间序列预测、事件检测、异常检测等。

2) 数据清洗阶段。根据收集的权威白名单域名数据对 DNS 数据进行清理，排除明显的非恶意域名流量以减少数据处理规模。通过数据清洗可以有效去除异常值或异常事件，保证模型正常运行，提高数据质量，确保最终结果可靠，减轻模型负担，减少内存占用。因此，总体来说，有效的数据清洗有助于提高模型的准确性，并确保整个流程的有效运行。

3) 频域转换阶段。将 DNS 请求时间序列转换为频域序列，并对每一个请求进行 DFT，目的是提取流量的时间序列信息，然后将频域序列转换为功率谱密度（PSD）向量，该向量表征不同频率下周期通信的强度，使神经网络算法能顺利训练模型。首先，从信誉良好的主机名中筛选 DNS 查询的时间序列，这些主机名不可能参与恶意活动。然后，将过滤后的 DNS 查询的时间序列按时间进行聚合。数据处理通过对聚合的 DNS 查询进行 DFT 来产生 PSD 向量。

4) 数据归一化阶段。归一化使预处理的数据被限定在一定的范围内，如[0,1]或者[−1,1]，从而消除奇异样本数据导致的不良影响，对数据进行标准化处理，确保不同特征的数据在数值上具有相同的尺度，以加快计算速度并稳定神经网络训练结果。对 DFT 产生的频域表示的模进行对数变换，防止在神经网络的训练过程中由于数值不稳定性问题而导致的浮点溢出。数据归一化的目的是统一数据维度，归一化后，优化过程的范围变小，优化过程将变得平稳，从而更容易正确地收敛到最优解。

5) 神经网络训练与分类阶段。该阶段的输入数据包括可信 DNS 流量，其被划分为良性设备的流量和注入恶意通信设备的流量。合成标记的数据用于训练神经网络，该网络根据请求是否由恶意主机发出，对新观察到的 DNS 请求的时间序列进行分类，神经网络模型根据进行 DNS 查询的设备是否良性对 PSD 向量进行分类。

1.1　时序处理

内部用户终端 H 在时间 T 对目标域名 D 发送的 DNS 查询 R 可以定义为

$$R =< H,D,T > \tag{1}$$

终端 H 可以设置为字符串 ID（如 ZJU06BA4F13）；目标域名 D 可以设置为 DNS 资源记录及资源类型；时间 T 可以设置为 Unix 时间戳，定义为从格林威治时间 1970 年 01 月 01 日 00 时 00 分 00 秒起至现在的总毫秒数。时序处理后，DNS 查询记录示例如下

< ZJU06BA4F13,google.com,A,1699848514206 >

表示在时间 1699848514206（北京时间 2023-11-13 12:08:34）访问 Google.com 的 A 记录。

FDS-DL 的输入数据集是终端 H_i 在指定时间段内（T_s 开始到 T_e 结束）发送的 DNS 查询序列，定义为

$$I_{(Hi,T_s,Te)} = \{ R(H = H_i) \wedge (T_s \leqslant T < T_e) \} \tag{2}$$

通过这种方式，FDS-DL 可以根据一系列传出的 DNS 查询来处理和分类设备“ZJU06BA4F13”是否是恶意主机，示例如下

$I_{(H=\text{ZJU06BA4F13})}$: {
< zju1897,google.com,A,1699860076962 > ,
< zju1897,google.com,MX,1699860280995 > ,
< zju1897,google.com,AAAA,1699860346985 >
}

1.2　数据清洗

FDS-DL 的输入可能包含大量对可信域名的查询。因此数据清洗阶段的目的是从输入 DNS 请求序列中删除非恶意域名的查询数据，以减少后续阶段处理时间并提高准确性，数据清洗主要基于 Majestic、Cisco Umbrella、DomainRank、Alexa 等全球权威数据源的排名前 100 万网站域名数据，交叉去重形成一份权威域名清单[16]。如果终端用户访问上述的公开列表（如 Cisco Top 1M 列表），则认定访问的目标域名是可信的。

定义过滤函数 F，它如果输出一个真值则表示请求 R 应保留，过滤函数使用特殊参数 G 表示可信的全球权威域名数量。如果 G 设置为 500000，则在权威域名列表排名前 50 万的目标域名被视为可信的，并将其从输入中排除。使用式(3)定义的过滤函数 F 对域名请求序列 R 进行清洗。

$$I^F(H_i,T_s,Te)=\{R(H=H_i)\wedge(T_s\leqslant T<T_e)\wedge(F(R))\} \tag{3}$$

DNS查询计数将前期清洗过的DNS请求序列转换为与等长时间关联的DNS请求数量，计数会导致信息丢失，但得到的计数显著小于过滤后的输入，从而减少了处理时间，过滤输入的计数为：

$$T_{(Hi,T_s,Te)}=(t_1,t_2,\cdots,t_i,\cdots,t_N) \tag{4}$$

其中，t_i表示在第i个时间段中进行的总体DNS查询的数量，N表示时间段的总数。

FDS-DL可以配置为使用不同的N，较低的值将导致较短的时间间隔，提供较优的数据细粒度;较高的值将产生较长的时间段从而提供更高的效率，高N值将导致系统检查更多的时间段，从而使其对一段时间内发生的检测更敏感，但效率较低，因为输入量变得更大，示例如下

$$T_{(N=168)}:(t_1=0,t_2=5,t_3=0,\cdots,t_{167}=5)$$

该例使用一小时（3600 s）的时间段和一周（7天×24小时）的时间（N=168）。每小时每个主机在一周内请求的一系列DNS查询都会被计算到N=168个时间段中，这些时间段表示对未知主机发出的DNS请求的数量。

1.3　频域转换

利用DFT对DNS请求数据在时间周期T内进行频域转换，表示为

$$X_{(T,k)}^{\mathrm{DFT}}=\sum_{n=0}^{N-1}t_n e^{-\frac{2\pi i}{N}kn}\ (k=0,1,\cdots,N-1) \tag{5}$$

PSD向量是周期性检测任务的常见数据表示形式，可定义为

$$V_i=\left\| X_{(T,k)}^{\mathrm{DFT}}\right\|^2 \tag{6}$$

$$P_{(T)}=\left(V_0,V_1,\cdots,V_{\frac{N-1}{2}-1}\right) \tag{7}$$

PSD向量定义为给定输入的DFT系数的幅值平方，PSD向量用于估计时间序列信号的频谱密度，这使神经网络分类器能够更容易地识别发生重复行为的时间。

假设时间周期数为一周（7天×24小时），则N=168，PSD值为

$$P_{(T,N=128)}:(V_0=0.326,V_1=0.157,\cdots,V_{82}=0.09)$$

因此，PSD向量可以被认为是特定时间频率下事件发生率的强度。

1.4　数据归一化

PSD向量的值是无界的并且具有不同的大小，因此在神经网络训练和分类之前必须对它们进行归一化处理，PSD向量的归一化将每个频率的幅度缩放到[0,1]区间，该值与训练集中出现的原始值成比例。所得到的归一化PSD向量保证对于每个频率保持相同的能量分布，并保持一致的分类尺度。归一化的PSD向量表示为

$$P'_{(T)}=\frac{\lg(p_{(T)})}{\lg\max(p_{(T)})} \tag{8}$$

归一化可以消除不同维度数据之间的差异，本文通过对数函数归一化方法，将原始DNS流量特征值重新缩放到[0,1]区间。

1.5　神经网络训练与分类

FDS-DL使用Keras构建了一个具有卷积层、池化层、全连通层和Softmax层的CNN，如图2所示。选择Adam作为神经网络优化算法，选择二元交叉熵（binary_crossentropy）作为损失函数，并将准确度作为评估指标。在分类问题中，理想输出通常是概率较高的输出，偶尔取概率较低的输出，因此本文应用了Softmax方法。

Adam优化器用于最小化二进制交叉熵损失，在Keras的Adam优化器中核心参数如下

keras. optimizers. Adam(learning_rate=0.001, beta_1=0.9, beta_2=0.999, epsilon=1e-07, amsgrad=

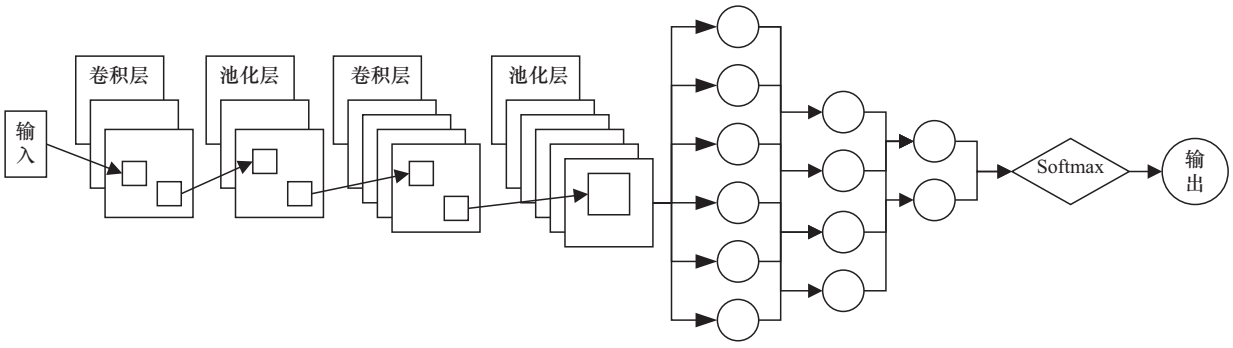

图2　CNN架构

False, weight_decay=None)

损失函数binary_crossentry如式(9)所示，其中$i \in [1,N_{output}]$相互独立。

$$\text{Loss} = -\frac{1}{N_{output}} \sum_{i=1}^{N_{output}} y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i) \quad (9)$$

本文采用ReLU函数，稀疏模型可以更好地挖掘相关特征并拟合训练数据，与其他激活函数相比，ReLU更具表现力，尤其是在深度网络中，对于非线性函数，ReLU不存在梯度消失问题，因为非负区间的梯度是恒定的，这使神经网络模型的收敛速度保持在稳定状态。

经上述训练完成的CNN作为函数W被应用于归一化PSD以输出分类结果C，如式(10)所示。

$$C = W(\theta, P'_{(T)}) \quad (10)$$

C是一个在[0,1]区间的连续值，其数值越高表示输入的一系列DNS域名查询由恶意程序发送请求的可能性更高。

2　模型评估

本文所需的实验数据来自浙江大学校园网络，通过部署软件定义网络（SDN）交换机将DNS流量镜像到存储节点，为了便于模型评估对比，将DNS数据包格式化为元数据，每天的元数据文件大约为400 GB，提取部分流量用于测试，以检测和分类正常流量和恶意流量，然后导出结果，观察模型检测各种类型流量的效率，找出可能存在的问题，并对模型和算法进行优化和校正，直到模型检测的准确性达到更好的结果。

本文将FDS-DL与其他主流的检测方法进行了比较，结果如表1所示。表1中，精度和F1_score是相应检测结果的平均值，对比方法的检测结果数据来自文献[12-15]。

虽然LSTM.MI和Various CNN方法的检测精度、F1_score可以达到94.97%、94.68%和96.82%、97.42%，但对于未参与模型训练的不同类型的恶意域名可转移性较差。FANCI和BotDigger方法基于多维低级特征，虽然它们的检测精度和F1_score超过97.7%，但它们需要大量的特征，并且很难预处理。这些方法对不同的机器学习算法具有高度的敏感性，并且用不同的机器学习算法处理相同的特征数据所获得的检测结果也不同。此外，同一算法的参数调整也依赖于人工先验知识。检测结果的鲁棒性较差，实时检测效果存疑。Deepdom方法通过关系图对提取的DNS流量特征进行分析，达到域名检测和识别的目的，尽管实现了较高的精度、F1_score和良好的鲁棒性，并且可以捕获域名之间丰富的关联信息，然而这类方法需要一个精确标记的大规模特征数据集来训练模型，这增加了模型的计算成本和运行压力，并导致了较大的检测时延。

表1　不同方法检测结果比较

模型	精度	F1_score
LSTM.MI	94.97%	94.68%
Various CNN	96.82%	97.42%
FANCI	98.86%	98.59%
BotDigger	97.87%	97.72%
Deepdom	93.86%	91.16%
FDS-DL	99.23%	99.17%

与其他检测方法相比，本文提出的FDS-DL不需要大量多样的特征和引入各种机器学习算法来分析特征，也不需要分析恶意样本，提高了检测结果的鲁棒性，减少了资源开销。FDS-DL不需要使用大量标记的数据集来构建地图类模型。与其他检测方法相比，FDS-DL在不存在模型过度拟合问题的情况下，对未知恶意域名具有更高的精度和F1_score，与基于先验知识的恶意域名检测方法相比，在保证算法便捷性的基础上，可以快速准确地检测出各种恶意域，提高了攻击者逃避检测的难度。综上，与其他方法相比，FDS-DL方法具有较高的精度和F1_score。

3　结束语

本文提出了一种恶意DNS流量检测方法FDS-DL，同时结合了频域特征聚合分析和神经网络算法对DNS流量进行分类。与主流检测方法进行了比较，实验结果表明，FDS-DL在检测精度和F1_score方面优于其他方法，具有较好的稳健性和能效性能。

参考文献：

[1] CrowdStrike. 2023 Global Threat Report[R].2023

[2] International Data Corporation. 2022 Global DNS Threat Report [R].2022.

[3] GRILL M, NIKOLAEV I, VALEROS V, et al. Detecting DGA malware using NetFlow[C]//Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management. Piscataway: IEEE

Press, 2015: 1304-1309.

[4] GAO J, ZHAO W H, ZHANG X, et al. MRI analysis of the ISOBAR TTL internal fixation system for the dynamic fixation of intervertebral discs: a comparison with rigid internal fixation[J]. Journal of Orthopaedic Surgery and Research, 2014, 9(1): 43.

[5] SCHÜPPEN S, TEUBERT D, HERRMANN P, et al. FANCI: feature-based automated NXDomain classification and intelligence[C]//Proceedings of the 27th USENIX Security Symposium. Berkeley: USENIX Association, 2018: 1165-1181.

[6] CASINO F, LYKOUSAS N, HOMOLIAK I, et al. Intercepting hail hydra: real-time detection of algorithmically generated domains[J]. Journal of Network and Computer Applications, 2021, 190: 103135.

[7] ALAEIYAN M, PARSA S, P V, et al. Detection of algorithmically-generated domains: an adversarial machine learning approach[J]. Computer Communications, 2020, 160: 661-673.

[8] ZHANG H , GHARAIBEH M , THANASOULAS S, et al. BotDigger: detecting DGA bots in a single network[C]//Proceedings of the Traffic Monitoring and Analysis. Berlin: Springer, 2016 :1-8.

[9] TRAN H, NGUYEN A, VO P, et al. DNS graph mining for malicious domain detection[C]//Proceedings of the 2017 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2017: 4680-4685.

[10] PENG C , YUN X , ZHANG Y ,et al. MalShoot: shooting malicious domains through graph embedding on passive DNS data[C]//Proceedings of the Collaborative Computing: Networking, Applications and Worksharing. Berlin: Springer, 2019: 488-503.

[11] YIN L H, LUO X, ZHU C S, et al. ConnSpoiler: disrupting C&C communication of IoT-based botnet through fast detection of anomalous domain queries[J]. IEEE Transactions on Industrial Informatics, 2020, 16 (2): 1373-1384.

[12] SUN X Q, WANG Z L, YANG J H, et al. Deepdom: Malicious domain detection with scalable and heterogeneous graph convolutional networks[J]. Computers & Security, 2020, 99: 102057.

[13] WOODBRIDGE J, ANDERSON H S, AHUJA A, et al. Predicting domain generation algorithms with long short-term memory networks[J]. arXiv Preprint, arXiv: 1611.00791, 2016.

[14] TRAN D, MAC H, TONG V, et al. A LSTM based framework for handling multiclass imbalance in DGA botnet detection[J]. Neurocomputing, 2018, 275: 2401-2413.

[15] VINAYAKUMAR R, SOMAN K P, POORNACHANDRAN P, et al. Evaluating deep learning approaches to characterize and classify the DGAs at scale[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(3): 1265-1276.

[16] STÉPHANE C, BLAKE S. A stable and open method for ranking domains[C]//Proceedings of the Internet Measurement Conference). New York: ACM Press, 2019: 1-7.

[作者简介]

单康康（1984– ），男，浙江东阳人，浙江大学高级工程师，主要研究方向为计算机体系结构、计算机网络安全、服务器系统研究等。

袁书宏（1974– ），女，四川达州人，浙江大学高级工程师，主要研究方向为网络信息安全、下一代网络技术。

陈文智（1969– ），男，广西田东人，博士，浙江大学教授、博士生导师，主要研究方向为嵌入式实时系统、分布式计算、虚拟化技术与可信计算等。

王志波（1984– ），男，浙江杭州人，博士，浙江大学教授、博士生导师，主要研究方向为人工智能安全、数据安全与隐私保护、边缘智能与安全等。