计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202440496
61(10): 2417−2432, 2024

基于数据集蜜点的抗损毁数据标识技术

李浩波[1] 李默涵[1] 陈 鹏[1] 孙彦斌[1,2] 田志宏[1,2]
[1](广州大学网络空间安全学院 广州 510006)
[2](广东省工业控制系统攻防对抗重点实验室 广州 510006)
(lhb@e.gzhu.edu.cn)

A Corruption-resistant Data Identification Technology Based on Dataset Honeypoint

Li Haobo[1], Li Mohan[1], Chen Peng[1], Sun Yanbin[1,2], and Tian Zhihong[1,2]
[1](*Cyberspace Institute of Advanced Technology*, *Guangzhou University*, *Guangzhou* 510006)
[2](*Guangdong Provincial Key Laboratory of Industrial Control System Security*, *Guangzhou* 510006)

Abstract Data identification is a prerequisite for achieving precise data governance, effectively ensuring the security of data elements during cross-domain transfer. Currently, there are methods for generating identifiers for individual data, but as the scale of data continues to expand, identifiers at the data level cannot be directly applied to the dataset level. This also introduces issues of identifiers being "easily damaged" and "difficult to embed". To effectively address these issues, we adopt the design concept of network honeypoint from the "guardian" model proposed by academician Fang Binxing. Utilizing the idea of deception defense, we propose an anti-damage data identification technology based on dataset honeypoint for cross-domain data transfer scenarios, and design a complete method for generating and embedding dataset honeypoints. First, for cross-domain data transfer scenarios, dataset honeypoints are designed. By enhancing the concealment of dataset honeypoints and increasing their redundancy, the issue of identifiers being "easily damaged" is addressed. Second, by ensuring that the form of dataset honeypoint is indistinguishable from real data, the issue of identifiers being "difficult to embed" is resolved. Finally, experiments conducted on both image and encrypted text data modalities demonstrate that dataset honeypoints possess high anti-damage capability, high robustness, and low performance overhead.

Key words data cross-domain transfer; data identification; dataset honeypoint; deception defense; damage resistance; embedding

摘 要 数据标识是实现数据精准监管的前提条件,有效地保证了数据要素在跨域流转过程中的安全.当前已有针对单一数据的标识生成方法,但是随着数据规模的不断扩大,数据层面的数据标识无法直接应

收稿日期: 2024-06-05; **修回日期**: 2024-07-16
基金项目: 国家重点研发计划项目(2021YFB3101704); 国家自然科学基金项目(62372126, 62272119, 62072130, U20B2046); 广东省基础与应用基础研究基金项目(2023A1515030142); 广州市基础与应用基础研究项目(2024A04J9969); 广州大学校内科研项目(YJ2023047); 广东省高校珠江学者资助计划项目(2019); 广东省高等教育创新群体项目(2020KCXTD007); 中国工程院战略研究与咨询项目(2023-JB-13)
This work was supported by the National Key Research and Development Program of China (2021YFB3101704), the National Natural Science Foundation of China (62372126, 62272119, 62072130, U20B2046), the Guangdong Basic and Applied Basic Research Foundation (2023A1515030142), the Guangzhou Basic and Applied Basic Research Foundation (2024A04J9969), the Guangzhou University Project (YJ2023047), Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2019), the Guangdong Higher Education Innovation Group Project (2020KCXTD007), and the Strategic Research and Consulting Project of the Chinese Academy of Engineering (2023-JB-13).
通信作者: 李默涵(limohan@gzhu.edu.cn)

用到数据集层面,并且会带来标识“易损毁”和标识“难嵌入”的问题.为有效解决上述问题,通过沿用方滨兴院士提出的“护卫”模式中网络蜜点的设计理念,借助欺骗防御的思想提出数据跨域流转场景下基于数据集蜜点的抗损毁数据标识技术,设计并形成一套完整的数据集蜜点生成和嵌入方法.首先,针对数据跨域流转场景设计了数据集蜜点,并通过增强数据集蜜点的隐蔽性和增加数据集蜜点冗余的方式解决标识“易损毁”的问题.其次,通过保证数据集蜜点形态与真实数据密不可分,解决标识“难嵌入”的问题.最后,通过在图像和加密文本2个数据模态下进行实验,验证了数据集蜜点具备高抗损毁、高鲁棒和低性能开销的特性.

关键词 数据跨域流转;数据标识;数据集蜜点;欺骗防御;抗损毁;嵌入

中图法分类号 TP391

党的十九届四中全会指出,要“健全劳动、资本、土地、知识、技术、管理、数据等生产要素由市场评价贡献、按贡献决定报酬的机制.”国家高度重视数据对经济社会发展的促进作用,为充分发挥数据要素乘数效应,《“数据要素×”三年行动计划》等一系列战略规划已陆续出台.由此可见,数据作为一种新型生产要素对国家和社会的全方位发展发挥着至关重要的作用.随着全球化和数字化的不断发展,在互联网技术快速发展的今天,数据已成为不可或缺的重要资源,医疗、金融、交通、政府等各行各业均需要数据作为支撑和保障,数据已快速融入生产、分配、流转和消费各个环节,以确保行业秩序的平稳推进.

一方面,数据流转具有必要性.数据在发挥其作用的过程中,涉及数据产生、数据分析、数据审计和数据应用等多个阶段[1].在完整的数据生命周期中,数据并非静态且一成不变的,数据流转是数据全生命周期中必不可少的过程.另一方面,流转中的数据具备更高的价值.“动起来”的数据可以更好地应用于开展业务研究和探索,促进行业内的相关业务快速发展.同时,数据的流转也可以促进行业间的互通和联动,促进跨行业协作以提高行业间的效率.全方位的数据互通和协同是数字治理的基础,保证数据依法有序自由流转至关重要,因此,需要开展针对数据安全相关的研究工作,以实现对数据的精准管控.

数据标识是实现数据精准管控的前提和首要条件,而数据精准管控是实现安全数据流转的基础.数据标识是一种唯一标注和识别单一数据元素或数据集的技术,其主要作用是确保数据的唯一性、可识别性和可追溯性,以支持数据交换、数据共享和数据集成等数据流转方式[2-3],从而为数据全生命周期的监督和安全管控提供支持和保障,数据标识现已被广泛应用于数据跨域流转领域.

高质量的数据标识需要满足4个性质,具体包括:1)便于确权.即数据标识应能追溯到数据的创建者或所有者,便于追踪数据来源以确定所有权的归属.2)易于解析.即数据标识应能恢复数据的属性信息,易于对数据信息进行准确解析和提取.3)抗损毁.即数据标识应具备不易被损毁和破坏的能力,抵抗在传输或存储过程中由人为因素带来的破坏.4)易于嵌入.即数据标识应具备轻易嵌入数据的能力,在对原始数据少做修改的基础上完成与数据的耦合.

当前已经有一些关于数据标识的研究工作,但当前方法更关注针对单一数据产生标识.根据国家互联网信息办公室发布的《个人信息和重要数据出境安全评估办法(征求意见稿)》的规定,50万条以上的个人数据可以构成重要数据.类似的情况也能够类比到其他类型的数据要素场景中.如果重要数据是一个包含大规模数据的数据集合,那么如何以数据集整体为目标,从数据集层面生成和嵌入数据标识,也是一个亟待解决的问题.

目前,物联网标识技术和数据水印技术都可以被用于生成跨域流转[4]过程所需要的数据标识.两者都能够实现针对单一数据的标识生成,可以满足数据标识便于确权和易于解析的性质.但是随着数据规模的不断扩大,当数据集中的数据达到一定的量级时,对数据集中的每条数据都生成并嵌入标识会带来大量的时间开销和资源消耗.这不利于海量数据场景下数据的高效快速跨域流转,最终影响数据要素充分发挥其作用价值.如果从数据集中选取部分数据,并利用现有方法进行单一数据标识生成会导致标识“易损毁”的安全性问题,这是因为数据层面的标识在数据跨域流转的过程中可能遭受恶意攻击者的刻意损毁,通过对识别到的标识进行丢弃、破坏或篡改,破坏数据标识的可用性从而影响标识的确权和解析,并最终给数据本身带来安全威胁.并且由于数据标识的产生在时间维度上落后于数据的产

生，数据标识很难以“无感”的方式与数据耦合，换言之，数据标识先天具备“难嵌入”的特点. 目前，针对标识“易损毁”和“难嵌入”的问题，尚无有效解决办法.

为此，本文沿用方滨兴院士提出的“护卫”模式的理念，结合网络蜜点[5]的设计思路提出基于数据集蜜点的抗损毁数据标识技术，基本思想如图 1 所示，通过对数据集属性信息编码形成数据集蜜点并嵌入数据集合以混淆攻击者对数据集蜜点的识别，旨在解决数据集层面的标识生成问题.



Fig. 1 Basic idea of dataset honeypoint

图 1 数据集蜜点基本思想

本文提出的基于数据集蜜点的标识方法与现有标识方法的特性对比如表 1 所示. 其中，易流转维度表示标识是否适用于数据跨域流转场景；轻量级维度表示标识生成和解析所需要的计算资源少，可以兼顾效率和灵活性；易部署维度表示标识生成方法实现难度小，能够在对现有系统仅做少量更改的情况下完成配置；抗损毁维度表示标识在流转过程中不易被破坏和损毁；易嵌入维度表示标识在数据产生之后仍能较容易地实现与数据的耦合；信息量维度表示标识能够携带的数据信息量；护卫模式维度表示标识生成方法是否属于护卫模式.

Table 1 Features Comparison of Different Types of Identification Methods

表 1 不同类型的标识方法的特性对比

标识方法	特性						
	易流转	轻量级	易部署	抗损毁	易嵌入	信息量	护卫模式
物联网标识	是	否	否	是	是	高	否
数字水印	是	是	是	部分	否	中/低	否
数据集蜜点（本文）	是	是	是	是	是	高	是

首先，本文方法通过提升标识隐蔽性和增加标识冗余来解决数据标识“易损毁”的问题. 具体来说，借助欺骗防御[6]的思想设计用作数据标识的数据集蜜点以欺骗攻击者，一方面，数据集蜜点能够实现数据集属性信息的记录，满足数据标注和识别的要求，另一方面，基于提取到的数据集属性信息生成在形式上与普通数据密不可分的数据集蜜点，以增强隐蔽性，提高潜在攻击者对数据集蜜点的识别难度. 一旦攻击者丧失准确甄别数据集蜜点的能力，则无法对数据集蜜点采取篡改、破坏等恶意行为，由此增强数据集蜜点的抗损毁能力，强化数据标识的安全性和可靠性. 此外，通过设置数据集蜜点的多份冗余，保证只要数据集中存在蜜点的 1 个备份就能够完整地解析数据集属性信息，并有效甄别标识信息是否被攻击者恶意篡改为虚假信息. 与此同时，数据集蜜点的多份冗余也能够很好地解决数据自然演进过程中导致的标识腐败过期的问题，避免因个别标识自然损坏导致数据标识整体不可用的情况，确保标识的有效性和可靠性，进一步提升数据标识的抗损毁能力. 确保在人为因素和自然因素导致的损毁场景中，数据标识在数据跨域流转的全过程中始终可用，实现敏感数据跨域流转的全方位精准监管.

其次，本文方法通过保证数据集蜜点形态与普通数据密不可分来解决数据标识“难嵌入”的问题. 由于蜜点形态与数据集中的普通数据形态具有一致性，本文将数据集蜜点作为普通数据采用向数据集中插入普通数据的方式，以一定规律将数据集蜜点分布于数据集中，完成嵌入的过程，从而保证嵌入的便捷性. 在此基础上，本文进一步探索了一种适用于数据集蜜点的基于属性信息的空间位置生成算法，用于计算数据集蜜点分布. 通过空间生成算法产生的结果，将数据集蜜点嵌入数据集中.

综上所述，本文为解决当前数据集标识生成方法所面临的“易损毁”“难嵌入”的问题，旨在提升数据标识的抗损毁能力和与原始数据集的耦合程度. 具体来说，本文的主要贡献包括 4 个方面：

1）设计一套适用于数据跨域流转领域的基于数据集蜜点的数据标识生成和嵌入方法，以实现对数据集合跨域流转的精准监管.

2）提出基于数据集蜜点的数据标识生成技术，借助护卫模式和欺骗防御的思想，保证数据标识可

用性的同时增强数据标识的隐蔽性,从而提升数据标识的抗损毁能力.

3)提出基于数据集属性信息的空间位置生成算法,通过设置数据集蜜点的多份冗余,增强数据标识嵌入过程的随机性和不可预测性,从而提升数据标识和原始数据集的耦合程度.

4)通过在图像和加密文本2个数据模态上进行实验,验证了数据集蜜点作为数据标识所具有的抗损毁、高鲁棒以及低性能开销等特性.

1 相关工作

本节主要介绍当前相关工作,梳理当前研究重点.

1.1 标识记录

数据标识是数据资产管理的基本手段,是与客体数据安全相关的属性的格式化封装,是数据安全属性的信息载体.为解决标识如何准确进行信息记录的问题,增强数据标识所承载的属性信息的有效性,从而为数据跨域流转过程提供重要的信息依据,研究人员开展了广泛的研究工作.

文献[7]针对数据库系统,从多源角度出发,提出一种用于解决数据源标记和中间源标记问题的模型,通过存储数据注释及其中间结果用于实现数据查询.Ram等人[8]提出经典的W7数据溯源模型,用来记录数据来源以及生命周期内处理、流转等关键过程,设计〈Where, Why, How, When, Who, Which, What〉七元组,形成可以正确表征数据的信息,以标记数据并跟踪影响数据的安全事件.文献[9]进一步阐述了W7模型中7个元素之间的相互关联关系,并证明W7模型具备足够的通用性和扩展性.文献[10]针对科学数据提出通用元数据设计规范化模型,元数据的本质是“关于数据的数据”,通过元数据描述信息资源或数据对象,实现对信息资源的识别及其使用过程的追踪.在W7模型的基础上,文献[11]进一步扩展了所记录数据的精细程度,补充Reference, Result, Remark三个元素构成W7-R3模型.

1.2 标识生成

1.2.1 物联网标识体系

随着物联网技术的发展和应用延伸,物联网所覆盖的领域逐渐拓宽,物联网标识为实现信息交换和数据流转提供了重要支持.当前物联网领域的标识主要分为网络标识和对象标识[12],特点对比如表2所示.

网络标识以Handl[13]和OID[14]为代表,通过结合标准的网络标识代码和用户特定的标识代码实现最终标识的生成.对象标识以GS1[15]和EPC[16]为代表,主要用于对商业化产品进行唯一标识编码,通过产品代码标准实现对产品的标识生成.作为我国自主研发的物联网标识解析体系,Ecode体系被应用于国家新基建基础设施建设中,还可应用于国家信息安全,诸如数字货币等众多领域.可以实现将物联网中抽象的实体以及虚拟化的对象采用唯一标识进行表示,并将编码后的信息嵌入载体,以便进行精准的识别、追踪和管理.通过采用安全性高的密码算法,实现安全可靠的标识设计.

Table 2 Features Comparison of IoT Identification Systems

表2 物联网标识体系特点对比

标识技术	应用场景	特点	不足
Handle	数字对象	持久灵活、分布式架构	管理维护成本高
OID	不区分	应用场景广泛、兼容性强	管理和使用复杂
GS1	商品流通	高效解析、可扩展性强	难以应对数据量快速增长的情况
EPC	实体对象	信息精确、自动化程度高	技术要求高、标识产生成本高
Ecode	实体对象	安全性高、兼容性强	用户体量小、技术和成本要求高

1.2.2 数字水印

数字水印是一种信息隐藏技术,将数字信号,如图像、文字、符号等一切可以作为标记的信息与原始数据紧密结合并隐藏其中,用于保护数字内容的版权和完整性,是一种广义上的数据标识,被广泛应用于图像和音视频领域.

Google Research的Luo等人[17]利用深度学习框架实现端到端可训练的视频水印方案,并通过水印检测器识别视频片段中包含水印的帧.韩国首尔大学的Yoo等人[18]提出基于自然语言处理文本特征的水印框架,在不影响语义质量的基础上,提升水印的鲁棒性和有效性,以保护自然语言版权.文献[19]提出一种用于大语言模型(large language model, LLM)的语义不变水印方法实现攻击鲁棒性和安全鲁棒性的权衡.

1.3 欺骗防御

随着网络的快速发展,作为网络安全策略之一的欺骗防御(deception defense)[20]逐渐进入大众的视野,欺骗防御旨在通过伪装、模拟真实网络服务或网络资产欺骗攻击者,增加攻击者识别真实目标的难度,从而实现对真实目标的保护.现如今,欺骗防御被越来越多的安全人员用于防御网络安全攻击.

“护卫模式”[4]作为一种主动防御模式，侧重于从全局视角发现并捕获攻击，支持全方位综合分析并保护目标系统，通过全局联动将所有信息汇聚起来. 在“护卫模式”下，网络蜜点主要用于全方位、多层次地主动感知被保护系统面临的安全威胁，通过对网络中的关键资源进行仿真实现蜜点的构建. 在部署了多类型蜜点系统中，攻击者往往需要花费更多的时间，采用更多的技术手段甄别真实攻击目标，区分蜜点和真正的系统资源，分散攻击者注意力的同时增加攻击者的攻击成本，进而降低攻击成功率.

欺骗防御的核心思想是利用 Honey-X 技术[21]，如 Honeypot, Honeytoken[22], Honeynet[23] 等形式上与真实资源不易区分的欺骗资源，干扰攻击者对真实目标的甄别，从而增加攻击者的攻击难度和成本，降低对真实目标的威胁. 通过对被保护系统的网络服务进行仿真，模拟服务端的应用程序或模拟网络协议，提高 Honeypot 的真实性从而增强 Honeypot 的隐蔽性. 最终影响攻击者对目标系统的识别，达到防御攻击的效果. Honeytoken[22]（也称为 Honeyfile），可以实现监视文件访问的功能. Honeytoken 需要模拟真实文件的文件名和类型，使其看起来与真实文件相似，同时将 Honeytoken 放置在系统中与真实文件类似的位置，增加攻击者的信任度，并通过对文件大小以及创建时间等文件属性的设定提高 Honeyfile 的隐蔽性.

由此可见，基于欺骗防御的 Honey-X 技术可以很好地用于混淆攻击者，增加攻击者对真实目标的识别难度，实现对真实目标的保护. 而本文将欺骗防御的思想应用于数据跨域流转领域，通过设计数据集蜜点实现对数据的标注和识别，使之与真实数据在形式上密不可分，混淆攻击者对数据集蜜点的识别从而提高数据标识在攻击者视角下的隐蔽性，确保数据集蜜点难于甄别，从而提高抗损毁能力，实现高抗损毁性的数据标识生成.

2 威胁模型和设计目标

本节主要对威胁模型和设计数据集蜜点的目标展开介绍.

2.1 威胁模型

本文重点关注数据标识的研究，即所产生的标识主要用于数据跨域流转领域对数据要素进行标注和识别，并不考虑物联网场景下针对物联网对象进行标识并追踪. 此外本文聚焦于数据集层面进行标识生成和嵌入的研究.

本文对攻击者的能力作出 3 点设定：

1）攻击者并没有操控数据集属性信息提取和记录的能力，这就意味着从原始数据集中所提取的数据集信息都是真实有效的.

2）攻击者没有决定数据集蜜点嵌入位置和操控嵌入过程的能力，这就意味着构建的数据集蜜点均能够按照既定结果准确嵌入数据集中.

3）攻击者能够在数据跨域流转过程中主动对数据标识进行识别并刻意损毁，即攻击者具备丢弃、破坏或将标识篡改为虚假信息的能力.

2.2 设计目标

本文为解决标识“易损毁”和标识“难嵌入”的问题，提出基于数据集蜜点的抗损毁数据标识技术，旨在提高数据集层面标识的抗损毁能力和耦合能力. 本文将数据标识“易损毁”的问题通过增强数据标识的隐蔽性解决，并通过设置数据集蜜点的多份冗余，解决数据标识“难嵌入”的问题，从而形成一套完整的适用于数据跨域流转领域的数据标识生成和嵌入方法.

本文提出的数据集蜜点作为一种新颖的数据标识，满足高质量数据标识所需要具备的 4 个重要性质：

1）便于确权. 数据集蜜点作为一种数据集层面的数据标识，与单一数据层面的数据标识相同，可以记录数据集包括创建者在内的多维度的属性信息，从而通过数据集蜜点中所携带的信息追踪数据来源以确定所有权的归属，从而确保便于确权.

2）易于解析. 数据集蜜点所携带的属性信息可以在消耗少量时间和资源成本的情况下完成提取，解析过程简单明确且不产生额外的开销，从而确保易于解析.

3）抗损毁. 作为数据标识的数据集蜜点在形式上与普通数据密不可分，此种设计方式能够分散攻击者注意力的同时增加攻击成本，使得攻击者难以对数据集蜜点进行准确甄别，从而提高数据集蜜点在攻击者视角下的隐蔽性以及确保抗损毁.

4）易于嵌入. 因数据集蜜点与普通数据形态上具备一致性，可以采用向数据集中插入普通数据的方式完成数据集蜜点的嵌入，从而确保易于嵌入.

3 数据集蜜点生成和嵌入方法

本节主要介绍适用于数据跨域流转领域的基于数据集蜜点的抗损毁数据标识技术的实现方式，介绍数据集蜜点的生成和嵌入的具体流程.

3.1 方法概述

本节设计数据集蜜点的生成和嵌入方法，该方法是一套完整的适用于数据跨域流转的数据集层面的标识生成和嵌入方法. 通常情况下，针对数据集的标识生成与针对单一数据的标识生成方法并不相同，单一数据层面的标识方法难以有效迁移至数据集层面，且数据层面的标识方法所产生的标识在形式上与数据本身存在较大差别，攻击者可以较轻易地发现并恶意破坏标识，从而影响对标识信息的解析和对数据所有权的确定. 因此，本文借助欺骗防御的思想设计了数据集蜜点用于对数据集属性信息进行记录，并参考仿照普通数据的结构进行数据集蜜点的生成，保证两者形式上的相似性，混淆攻击者对数据集蜜点的识别从而降低数据集蜜点被恶意破坏和损毁的风险.

本文所提出的数据集蜜点的总体实现流程如图 2 所示，包括 3 个关键部分，即数据集属性提取、数据集蜜点生成以及数据集蜜点嵌入. 数据集属性信息通常由数据所有者提供，不需进行额外的操作以获取数据集信息. 为确保数据集蜜点实现流程上的完整性和逻辑上的连贯性，本文将数据集属性提取作为一个单独的模块进行分析.

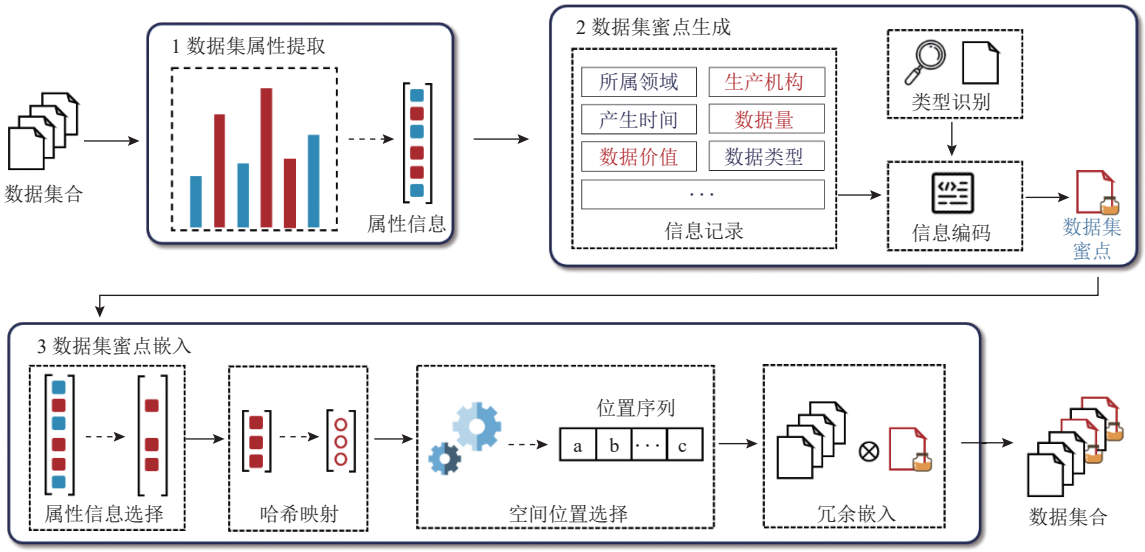


Fig. 2 Generation and embedding process of dataset honeypoint

图 2 数据集蜜点生成和嵌入流程

针对包含同一类型的数据集，首先收集并统计原始数据集的属性信息，包括数据量、产生时间等维度的信息形成属性信息向量，依据属性信息向量进行记录并形成用于后续生成数据集蜜点的信息载体. 需要指出的是，数据集属性提取模块进行全方位统计和分析的目的是为了获得可以用于描述数据集属性的全方位信息，以实现数据集蜜点便于确权的特性.

参考数据集中普通数据的类型和结构，将信息载体编码形成与普通数据形式上相近的数据集蜜点从而完成数据标识的生成.

经过数据集属性提取得到的属性信息向量在数据集蜜点嵌入阶段被首先用于关键信息选择，对被选择的关键信息进行哈希映射操作以用于空间位置的生成，关键特征的选择可以很好地降低用于哈希映射的信息维度，降低哈希映射过程的时间开销. 空间位置选择过程会得到位置序列，并将多份冗余的数据集蜜点按照位置序列插入数据集中从而实现标识嵌入.

3.2 数据集属性提取模块

数据集属性提取模块的主要目的是从数据集中统计并分析出数据集所携带的属性信息. 为方便描述，现将数据属性信息分类为定性属性信息和定量属性信息. 定性属性信息必须由数据所有者提供，包括产生者、所属领域等，用 V_x 表示. 而定量属性信息可以通过直接从数据所有者获取，或依据以下方法获得，用 V_1 表示. 为描述方便，现仅对数据量、数据集产生时间进行统计.

令 $D=\{d_1, d_2, \cdots, d_n\}$ 表示原始数据集合，获取每一条数据的信息，包括其创建时间 t 和数据大小 s. 首先统计本数据集合所包含的数据量 N，并以最晚数据产生时间作为数据集产生时间，即产生时间 $T=\max(t_1, t_2, \cdots, t_n)$.

定量属性信息统计完成后得到 V_l, 从定性属性信息提取数据所属领域 F、数据类型 C、生产机构 P 以及数据价值 S 等信息得到 V_x, 共同构成属性信息向量 $\boldsymbol{V}$=(F, P, C, N, S, T, …). 检查属性信息向量 $\boldsymbol{V}$ 的完整性, 要求其中前 6 个字段 v_1, …, v_6 不为空, 出于简单性的考虑, 完整性验证操作仅需对属性信息向量 $\boldsymbol{V}$ 前 6 个字段是否非空进行验证, 经过完整性验证的属性信息才可用于下一步数据集蜜点的生成.

3.3 数据集蜜点生成模块

数据集蜜点生成的主要目标是将数据集属性提取模块获得的属性信息向量编码形成数据集蜜点, 保证数据集蜜点与普通数据在形式上不易区分, 以解决数据标识“易损毁”的问题.

本模块将原有数据标识容易被攻击者破坏和损毁的问题, 采用标识欺骗的方式解决, 将数据集蜜点设计成普通数据增强数据集蜜点的隐蔽性从而提高数据集蜜点的抗损毁性质.

3.3.1 信息记录

信息记录的目的是将属性信息以格式化的形式进行存储并记录. 通过第 2 节威胁模型部分的介绍可以得知, 本文假设攻击者没有操控数据集属性信息提取和记录的能力, 即认为信息记录的过程不会产生来自外部的安全性问题.

本文采用列表的形式对属性信息进行记录和存储, 以满足属性信息的动态更新和异构化要求. 随着属性信息提取模块所获得信息向量的维度扩张或缩减, 采用列表化的属性信息记录能够很好地满足动态增减要求, 同时列表的异构化满足属性信息结构类型多样化的需要. 需要指出的是, 图 2 中信息记录阶段主要呈现数据集蜜点所承载和记录的数据集属性信息的样例.

3.3.2 类型识别

类型识别的主要目的是获取原始数据集的类型, 如图像、文本等, 以根据其不同类型的结构特点有针对性地生成数据集蜜点. 数据集蜜点旨达成与普通数据在结构形式上密不可分的欺骗目标, 类型识别阶段为数据集蜜点的构建提供结构信息.

本文选取具备复杂结构的文本类型数据, 以清晰呈现如何准确识别数据的结构信息并以此为依据生成数据集蜜点. 从文本类型数据中选择数据包文件进行分析, 说明如何识别并提取结构信息. 数据包具有物理层、数据链路层、网络层和传输层等多层嵌套结构. 通过逐层解析提取每一层的数据包信息, 例如以太网帧中所包含的 MAC 地址、IP 层所携带的数据包存活时间(time to live, TTL)、协议等, 以及 TCP/UDP 层的端口号、应用层的有效载荷 Payload 长度等. 源、目的 MAC 地址, 源、目的 IP 地址和源、目的端口需要成对出现以满足数据包发送过程的现实逻辑. 本文采用类似于 Bartos 等人[24] 的数学表示记录数据包的结构信息, 使用 f_i 表示第 i 个数据包的结构信息, m 表示数据包结构信息的维度, 所有数据包的结构信息由矩阵 $\boldsymbol{F}$ 表示, $\boldsymbol{F}$ 定义如式(1)所示, 其中, f_{ij} 表示第 i 个数据包第 j 个结构信息.

$$\boldsymbol{F}=(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m)=\begin{pmatrix} f_{11} & \cdots & f_{1k} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mk} \end{pmatrix}. \tag{1}$$

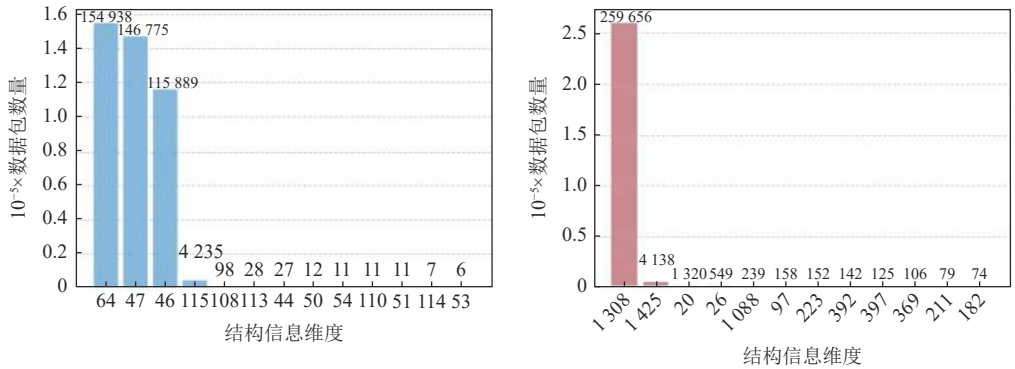
需要指出的是, 由于不同类型数据的结构差异性, 目前仅以数据包类型数据进行举例说明, 介绍其结构信息提取和记录方式.

3.3.3 信息编码

信息编码的目的是将承载了数据集属性信息的结构化数据构建形成数据集蜜点, 保证数据集蜜点具备标识数据功能的同时, 通过模拟普通数据的结构特征提升自身隐蔽性.

为承接类型识别阶段的过程, 本文仍以数据包文件为例进行信息编码阶段的介绍. 类型识别阶段为编码阶段提供数据的结构信息用于构成数据集蜜点. 获取结构信息矩阵 $\boldsymbol{F}$ 后, 对每一维度的数据包结构信息进行数量统计, 并使用直方图[25] 表示数据包的结构信息分布情况. 图 3 展示了 ISCXVPN2016 数据集其中一个数据包文件的 IP 层的 TTL 以及应用层的 Payload 长度的统计情况, 通过直方图可以直观看出信息矩阵 $\boldsymbol{F}$ 中每一维度结构信息的统计结果. 为增大数据集蜜点的真实性, 选择各层信息统计结果的最大值对应的内容作为构建虚假数据包的结构信息并据此进行数据包的层次化搭建. 经过上述操作便完成虚假数据包的构建, 使其具有与当前数据集最高的结构信息相似度, 从而保证虚假数据包与普通数据形式上不可区分, 完成数据集蜜点构建的第 1 步.

将记录数据集属性的结构化信息以有效载荷的方式与数据包结构耦合. 因数据集蜜点应满足便于解析的特点, 因此需要采用能够准确编码并恢复的方式完成属性信息向数据集蜜点的“隐写”. 在进行结构化信息隐写之前, 使用加密算法对数据集属性信息进行加密, 并按照 Payload 部分采用的字节串表示方式记录加密后的属性信息. 具体来说, 本文采用 AES 加密算法进行属性信息加密并使用加密块链接(cipher block chaining, CBC)[26] 模式进一步增强加密



(a) IP层TTL大小统计 (b) 应用层Payload长度统计

Fig. 3 Distribution of packet structure information

图 3 数据包结构信息分布

过程抵抗破解的能力, 以降低被解码的风险. 经过上述操作便完成数据包类型的数据集蜜点的创建, 使其所承载的属性信息以密文形式存储, 从而保证属性信息的安全性, 完成数据集蜜点创建的全部流程.

3.3.4 其他模态数据集蜜点生成举例

针对图像模态数据, 依据图像的像素大小和RGB属性等信息, 将3.2节中提取到的数据集属性信息编码形成与普通数据形式上密不可分的数据集蜜点. 生成的数据集蜜点中蕴含数据集的属性信息, 依据CIFAR-10数据集中的部分图像生成数据集蜜点, 具体结果如图4所示.



(a) 数据集蜜点

(b) 真实图像

Fig. 4 Dataset honeypoint of image modality

图 4 图像模态的数据集蜜点

3.4 数据集蜜点嵌入模块

数据集蜜点嵌入的主要目标是将数据集蜜点生成模块所构建的数据集蜜点, 采用预先设定的分布方法嵌入原始数据集中, 保证数据集中存在多份蜜点冗余的情况下增强数据集蜜点位置的随机性, 以解决数据标识“难嵌入”的问题.

本模块将原有数据的标识与数据本身耦合程度不高的问题, 采用数据插入的方式解决, 并按照特定规律与原始数据一同存放, 从而完成嵌入的过程.

3.4.1 属性信息选择

属性信息选择的目的是从数据集属性提取模块获得的属性信息向量中筛选出用于哈希映射的关键信息, 该做法在很大程度上降低了哈希映射过程的时间开销, 通过关键信息选择实现原有属性信息向量的维度压缩. 属性信息选择在全方位描述数据集整体情况的基础之上, 通过筛选的方式获得低维信息向量用作空间位置生成算法的种子, 既避免手动选择种子带来的操作上的复杂, 又可以基于不同数据集产生不同的种子, 兼顾时间成本的同时实现自动化生成.

本文将属性信息选择的问题形式化为可满足模理论(satisfiability modulo theories, SMT)问题, 目标是满足给定信息筛选条件的前提下找到关键信息. 二元变量s_i表示属性信息向量$\boldsymbol{V}$中的第i维信息v_i是否被选取, 如果v_i被选取则$s_i=1$, 否则$s_i=0$. 考虑到时间开销等实际条件, 关键信息应满足多个约束.

关键信息选择的类型约束形式化为式(2), 其中这2个不等式分别表示从属性信息向量中至少选择1个维度的定性属性信息作为关键信息, 以及至少有1个定量属性信息被选择.

$$\begin{aligned}&\sum_{i=1}^{len(V_x)} s_i>0,\ \forall v_i\in V_x,\\&\sum_{i=1}^{len(V_l)} s_i>0,\ \forall v_i\in V_l.\end{aligned}\tag{2}$$

关键信息选择的数量约束形式化为式(3), 从被筛选后的关键信息向量维度进行限制.

$$0<\sum_{v_i\in V} s_i<3.\tag{3}$$

关键信息选择的性能约束形式化为式(4), 其中τ_{delay}表示每一维度信息在哈希映射过程所带来的时间延迟, k表示最大时间延迟的阈值. 式(4)的约束使得所筛选的关键信息满足后续进行哈希映射的时间要求.

$$\sum_{i=1}^{len(V)} s_i \tau_{\text{delay}} \leqslant k,\ \ \forall v_i \in V_l. \tag{4}$$

通过对属性信息进行类型层面、数量层面和性能层面的约束限制，并使用 Z3 SMT 求解器[27]解决约束满足问题，从而实现关键属性信息的选择，得到可用于哈希映射的低维属性信息向量.

3.4.2 哈希映射

哈希映射的目的是将属性信息选择后的信息向量编码为空间位置选择阶段算法可以利用的种子，从而用于数据集蜜点待嵌入空间位置序列的生成.

因用于数据集蜜点待嵌入空间位置序列生成的种子需要具备高安全性和唯一性的要求，以保证不同属性信息向量编码后的结果难预测且各不相同. 且为满足简单性的要求，本文采用 SHA-256 哈希函数对低维属性信息进行编码，因 SHA-256 哈希函数计算哈希值的过程具备高安全性和抗碰撞性，可以很好地满足算法种子对于安全性和唯一性的需求，保证空间位置生成的不可预测性和唯一性. 经过 SHA-256 哈希函数的编码映射的结果在进制转化后即得到安全种子，上述整个过程可以形式化为式(5).

$$seed = HashMap(\boldsymbol{V}'), \tag{5}$$

其中 *seed* 表示生成的安全种子，*HashMap* 表示编码函数，$\boldsymbol{V}'$表示低维信息向量.

3.4.3 空间位置选择

空间位置选择的目的是在数据集所包含的空间范围内产生指定数量的位置用于数据集蜜点的插入，以完成数据集蜜点嵌入的过程，增强数据集蜜点位置的随机性，保证数据集中存在多份冗余，实现数据集蜜点与原始数据集的紧密耦合.

数据集蜜点的作用机制是在数据集流转过程中对数据集完整性、所有权等信息进行记录并基于数据集蜜点进行安全监控，数据集被接收后需要精确定位数据集蜜点的位置以实现数据集蜜点所记录内容的解析和恢复. 为满足空间位置生成的可逆性，本文提出一种基于数据集属性信息的空间位置生成算法，以实现空间位置的生成和恢复，同时保证所生成空间位置的难以预测性，如算法 1 所示.

算法 1. 基于数据集属性信息的空间位置生成算法.

输入：属性信息向量 $\boldsymbol{V}'$，空间位置数量 *N_pos*；

输出：空间位置序列 *SpaPos*.

① Initialize the parameters α, β, μ; /*初始化列表 *SpaPos* 用于存储空间位置*/

② $SpaPos = []$; /*将向量编码成字种子*/

③ $V_{str} = Turn_Vector_into_String(V')$;

④ $Pos_old = seed = HashMap(V')$;

⑤ for *index* from 0 to *num*−1 do /*递归生成新的空间位置 *Pos_new* */

⑥ $Pos_new = (\alpha \times Pos_old + \beta) \bmod \mu$; /*将 *Pos_new* 添加至 *SpaPos**/

⑦ update *SpaPos* with *Pos_new*;

⑧ $Pos_old \leftarrow Pos_new$;

⑨ *index*++;

⑩ end for

首先，初始化参数 α, β, μ，并且创建用于存储空间位置的列表 *SpaPos*. 其中 α, β 为空间位置生成算法的超参数，μ 表示最大生成范围的阈值，即支持数据集蜜点嵌入的空间位置上界. 随后将低维属性信息向量 $\boldsymbol{V}'$ 拼接构成一维字符串，并采用哈希映射方法得到种子 *seed*.

其次，在得到 *seed* 后，将其用于第 1 个空间位置 Pos_1 的计算，计算方式如式(6)所示.

$$Pos_1 = (\alpha \times seed + \beta) \bmod \mu. \tag{6}$$

随后，采用基于递归的方式计算得出后续所有空间位置，即新的数据集蜜点嵌入位置 *Pos_new* 的生成基于旧的位置 *Pos_old*，具体递归的计算方式为

$$Pos_new = (\alpha \times Pos_old + \beta) \bmod \mu. \tag{7}$$

最后，将新生成的数据集蜜点位置 *Pos_new* 加入空间位置序列 *SpaPos*，重复式(7)的递归计算方式，直至空间位置序列 *SpaPos* 的长度达到所需的空间位置数量 *N_pos*，需要指出的是，*N_pos* 可以被指定，在默认情况下 *N_pos* 的值等于数据集数量 *N* 的 1%. 通过基于属性向量的空间位置生成算法便实现了空间位置序列的生成.

3.4.4 冗余嵌入

冗余嵌入的目的是将数据集蜜点按照空间位置选择阶段生成的空间位置序列所指示的位置进行插入，实现对数据集蜜点和数据集的耦合，完成数据集蜜点嵌入模块的最后一步流程.

本文设计冗余嵌入的出发点在于设置数据集蜜点的多次备份，即通过插入多个数据集蜜点进一步提高数据集蜜点的安全性. 每个数据集蜜点均记录数据集属性信息，以保证即便数据集中某个或某几个数据集蜜点被丢失或被损毁的情况下，仍可以实现对数据集所有权等信息的确定. 进一步提升数据集蜜点的抗损毁性.

冗余嵌入算法的具体实现流程如算法 2 所示.

算法 2. 冗余嵌入算法.

输入: 原始数据集合 D, 空间位置序列 *SpaPos*, 数据集蜜点 M;

输出: 嵌入数据集蜜点的数据集 *D_hp*.

① *D_hp* = []; /*将集合列表化*/

② *M_list* = *List*(*M*), *D_list* = *List*(*D*), *L* = *len*(*D_list*) + *len*(*SpaPos*);

③ for each *pos* in *range*(*L*) do /*判断位置 *pos* 是否应嵌入数据集蜜点 *m**/

④ if *pos* in *SpaPos* then;

⑤ *m* = *M_list*[*SpaPos.index*(*pos*)];

⑥ *D_hp.append*(*m*);

⑦ end if

⑧ else

⑨ *D_hp.append*(*D_list.pop*(0));

⑩ *pos*++;

⑪ end for

首先, 对生成的数据集蜜点进行索引标注以便对每个数据集蜜点精确定位, 用 $M=\{m_1, \cdots, m_x\}$ 表示 x 个带有索引的数据集蜜点集合; 其次, 使用冗余嵌入算法遍历所有可能的空间位置, 并将数据集蜜点依次嵌入空间位置序列 *SpaPos* 指定的位置, 最终构成嵌入数据集蜜点的数据集合 *D_hp*.

在数据集完整的情况下, 可以依据 3.4.3 节中描述的基于属性信息的空间位置生成算法计算出数据集蜜点在数据集中嵌入的位置, 从而获取数据集蜜点用于解析以及确权, 并将数据集 *D_hp* 恢复成为初始状态. 如果数据集 *D_hp* 中的数据在跨域流转的过程中被部分丢弃, 则采取动态空间窗口的方式在其中寻找数据集蜜点, 以准确解析并确权. 空间窗口的大小随数据丢失率动态变化, 以保证数据蜜点的搜寻效率. 通过空间窗口中遍历所有的空间位置, 检测其中是否包含数据集蜜点, 实现数据集蜜点的初步定位并记录数据集蜜点的位置, 将该位置与 3.4.3 节中基于属性信息的空间位置生成算法的结果 *SpaPos* 进行比较, 从中定位到与数据集蜜点最相近的位置, 并基于数据集蜜点到当前位置和 *SpaPos* 序列元素之间的位置间隔, 进行剩余数据集蜜点位置的迭代搜寻.

4 实验验证

本节针对数据集蜜点开展实验验证, 在多模态场景中从隐蔽性、鲁棒性以及性能开销 3 个方面进行评估, 并对实验结果进行分析总结.

4.1 实验设置

4.1.1 数据集

本文分别从图像数据[28]和加密数据 2 个数据模态对数据集蜜点的隐蔽性、有效性、鲁棒性以及性能开销进行验证, 以 CIFAR-10 数据集和 ISCXVPN-2016 数据集作为代表, 对文本数据、图像数据以及加密数据在实验室环境中展开实验以模拟不同类型的数据集跨域流转的场景.

1)CIFAR-10 数据集[29]. 接近现实世界物体的彩色图像数据集, 该数据集包括 airplane, dog 以及 bird 在内的 10 种类别 RGB 图像. 本文选择该数据集模拟数据跨域传输领域的图像数据, 以验证数据集蜜点在图像数据领域的表现.

2)ISCXVPN2016 数据集[30]. 在 ISCX 中生成的现实世界流量数据集, 该数据集包含使用 Facebook, Skype 以及 Bittorrent 等服务获取的常规流量和 VPN 流量. 本文选择该数据集模拟数据跨域流转领域的加密数据, 以验证数据集蜜点在加密数据领域的表现.

4.1.2 评价指标

本文使用余弦相似度(*cosine similarity*)、欧氏距离(*Euclidean distance*)衡量数据集蜜点与原始数据的特征相似性, 具体计算公式如式(8)和式(9)所示, 其中 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 分别表示 2 个不同的向量.

$$cosine\ similarity(\boldsymbol{V},\boldsymbol{W}) = \frac{(\boldsymbol{V},\boldsymbol{W})}{\|\boldsymbol{V}\|\cdot\|\boldsymbol{W}\|}, \tag{8}$$

$$Euclidean\ distance(\boldsymbol{V},\boldsymbol{W}) = \sqrt{\sum_{i=1}^{n}(\boldsymbol{V}_i-\boldsymbol{W}_i)^2}. \tag{9}$$

本文使用峰值信噪比(peak signal-to-noise ratio, *PSNR*)和结构相似性(structural similarity, *SSIM*)衡量图像类型数据集蜜点与原始图像之间的差异性, 具体计算公式如式(10)(11)所示.

$$PSNR = 10\times\log\left(\frac{MAX^2}{MSE}\right), \tag{10}$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}, \tag{11}$$

其中, 式(10)中的 *MAX* 是图像的最大可能像素值, *MSE* 是均方误差. 式(11)中的 μ_x 和 μ_y 表示图像的平均值, σ_x^2 和 σ_y^2 表示图像的方差, σ_{xy} 是图像的协方差, C_1 和 C_2 是 2 个常数.

为验证数据集蜜点抵抗攻击的能力, 本文提出保留率(*Retain Rate*)这一指标, 计算数据集蜜点在遭受破坏后仍保留在数据集中的数量占总数的百分比, 从而衡量数据集蜜点的鲁棒性, 保留率的计算公式如式(12)所示.

$$Retain\ Rate=\frac{\sum_{i=1}^{len(D_{retain})} r_i}{len(SpaPos)},\quad r_i=\begin{cases}1, \text{ if } r_i\in SpaPos\\ 0, \text{ 其他}\end{cases},\tag{12}$$

其中，*SpaPos* 指的是空间位置序列，D_{retain} 指代遭破坏后的数据集合，r_i 表示空间位置，当 r_i 是序列 *SpaPos* 中的元素时其取值为 1.

4.2 隐蔽性验证

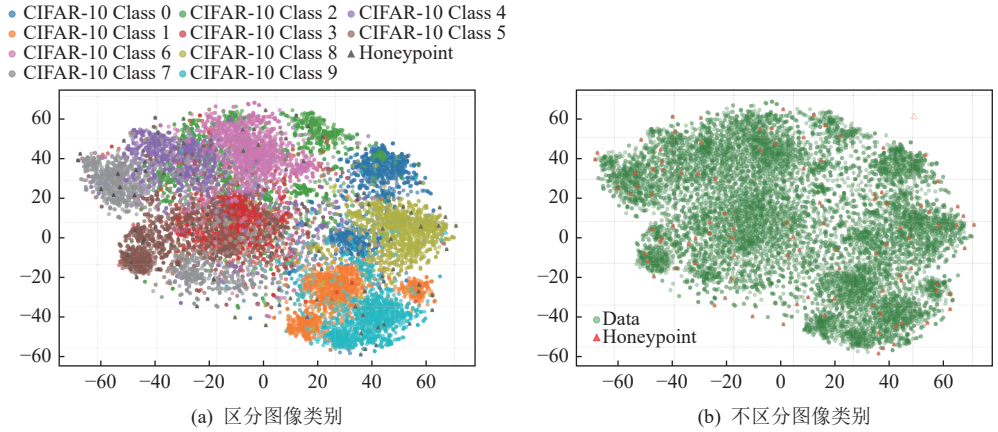
为有效验证数据集蜜点的隐蔽性，本文分别使用 CIFAR-10 数据集和 ISCXVPN2016 数据集开展实验.

4.2.1 图像数据模态

本文从特征维度和数据维度衡量展示图像数据模态中数据集蜜点的隐蔽性，特征维度采用特征可视化和客观衡量 2 个角度进行实验.

在特征可视化角度，依据 CIFAR-10 数据集的结构信息，在图像数据总量的基础上生成 1% 的数据集蜜点，并按照图 2 所示的嵌入流程完成数据集蜜点与 CIFAR-10 数据集的耦合. 本文采用 ResNet18[31] 网络模型对 CIFAR-10 数据集和数据集蜜点进行特征提取，通过移除 ResNet18 的最后一个全连接层，只使用网络结构中的卷积层和全局平均池化层提取图像的特征，采用 t-SNE[32](t-distributed stochastic neighbor embedding)降维方式对数据集进行可视化绘图. 以直观呈现数据集蜜点与图像数据集在二维空间中的分布情况.

图 5 呈现了降维后的数据集蜜点和原始数据特征的分布情况. 图 5(a) 在区分图像类别的条件下，将 CIFAR-10 数据集原有的 10 个类分别由 10 种彩色的圆形表示，使用黑色三角形表示数据集蜜点. 通过可视化结果可以直观看出数据集蜜点在每个类别中均有分布，且数据集蜜点与每个类别在特征空间中均具备相似性. 图 5(b)将整个数据集视为一个整体，使用绿色圆形表示全部数据，使用红色三角形表示数据集蜜点. 在不区分图像类别的条件下可以看出数据集蜜点在整体层面与整个数据集特征不可区分，从而说明图像数据领域数据集蜜点具备很高的隐蔽性.



(a) 区分图像类别　　(b) 不区分图像类别

Fig. 5 Dataset honeypoint distribution diagram based on CIFAR-10 dataset

图 5 基于 CIFAR-10 数据集的数据集蜜点分布图

在客观衡量角度，为进一步验证数据集蜜点的隐蔽性，本文使用 4.1.2 节列举的余弦相似度等 3 个维度的指标衡量数据集蜜点与 CIFAR-10 数据集中每个类别在特征空间中的差异性，并与像素值差分(pixel value differencing, PVD)[33] 算法进行对比.

同样地，采用 ResNet18 网络模型进行特征提取，并分别计算数据集蜜点和 PVD 算法所生成图像的特征与 CIFAR-10 数据集各类别特征在整体上的差异性. 表 3 衡量了 2 种方法生成的图像与 CIFAR-10 数据集 10 个类的差异的均值. 本实验生成数据总量 1% 的数据集蜜点，PVD 算法所生成的图像数量也占数据总量 1%. 从表 3 可以看出，数据集蜜点特征在余弦相似度、欧氏距离以及结构相似性方面 3 个指标衡量上均取得了不错的效果. 在绝大多数类别中，数据集蜜点与真实图像的特征相似度均高于 PVD 算法所生成图像与真实图像的相似度，从特征角度客观展示了本文所提出的数据集蜜点具备高隐蔽性.

为深入客观衡量数据集蜜点的隐蔽性，本文采用 *PSNR* 和 *SSIM* 指标来衡量数据集蜜点与其参照图像在数据空间中的差异性，同时与 PVD 算法所生成图像作为对照. 从表 4 可以看出，在 3 种比例下，数据集蜜点的 *PSNR* 和 *SSIM* 值均高于 PVD 算法所生成的图像. 相较于 PVD 算法，数据集蜜点的 *SSIM* 数值在 3 种比例下分别提升 5.07 个百分点、4.84 个百分

Table 3 Feature Dimension Concealment Verification Results Based on CIFAR-10 Dataset

表 3 基于 CIFAR-10 数据集的特征维度隐蔽性验证结果

算法	指标	CIFAR-10 数据集类别									
		类别 0	类别 1	类别 2	类别 3	类别 4	类别 5	类别 6	类别 7	类别 8	类别 9
数据集蜜点	余弦相似度	0.661 0	0.690 1	0.655 7	0.670 8	0.667 1	0.663 5	0.651 4	0.683 4	0.679 4	0.691 9
	欧氏距离	22.989 0	22.066 0	23.490 0	22.812 0	22.645 0	23.533 0	23.050 0	22.986 0	22.420 0	22.464 0
	结构相似性	0.019 3	0.009 8	0.007 6	0.002 6	0.002 5	0.000 8	0.002 7	0.007 5	0.021 3	0.018 4
PVD[32]	余弦相似度	0.657 2	0.704 2	0.648 9	0.666 9	0.656 0	0.660 6	0.648 3	0.677 1	0.703 6	0.705 7
	欧氏距离	22.676 0	21.896 0	23.965 0	22.865 0	23.204 0	23.970 0	23.301 0	23.610 0	21.886 0	22.414 0
	结构相似性	0.020 2	0.007 9	0.004 9	0.000 9	–0.001 1	–0.000 4	–0.001 4	0.003 4	0.020 5	0.014 0

Table 4 Data Dimension Concealment Verification Results Based on CIFAR-10 Dataset

表 4 基于 CIFAR-10 数据集的数据维度隐蔽性验证结果

算法	指标	生成比例		
		1%	5%	10%
数据集蜜点	*PSNR*	30.795 0	30.659 0	30.793 0
	SSIM	0.996 6	0.995 6	0.995 8
PVD	*PSNR*	29.394 0	29.317 0	29.411 0
	SSIM	0.948 5	0.949 6	0.950 6

点、4.75 个百分点, *PSNR* 分别提升 1.401, 1.342, 1.382.

为直观衡量数据集蜜点和 PVD 算法在图像数据模态所生成图像的隐蔽性效果, 本文分别针对表 4 的结果对 *PSNR* 和 *SSIM* 进行可视化绘图, 结果如图 6 所示. 说明本文方法生成的数据集蜜点与原始图像相似度更高, 展示出数据集蜜点良好的隐蔽性.

通过上述 3 个角度开展的隐蔽性验证, 验证了图像数据模态下数据集蜜点具备高隐蔽性, 从而说明数据集蜜点具备高抗损毁性.

4.2.2 加密数据模态

在 ISCXVPN2016 数据集中选取结构信息差异大的 Vpn_bittorrent.pcap 和 Facebook_audio3.pcapng 作为原始数据集. 分别依据 2 个文件所包含的数据包结构信息, 生成数据包总量 1% 的数据集蜜点并按照图 2 所示的数据集蜜点嵌入流程完成与数据集的耦合.

本文采用 Python 的 Scapy 库提取数据包和数据集蜜点的特征, 并采用 t-SNE 降维方式对特征进行可视化绘图, 以直观呈现数据集蜜点与加密数据集在二维空间中的分布情况.

如图 7 所示, 红色三角形表示数据集蜜点在二维空间中的分布情况, 蓝色圆形表示原始数据包在二维空间中的分布情况. 图 7(a)表明对于结构信息较为分散的 bittorrent 数据集, 本文所提出的数据集蜜点生成方式能够广泛地获取其不同的结构信息并依

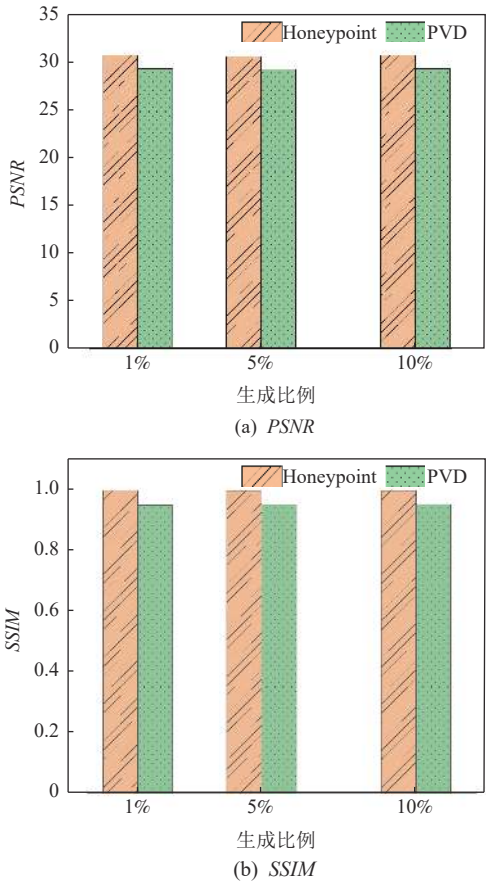
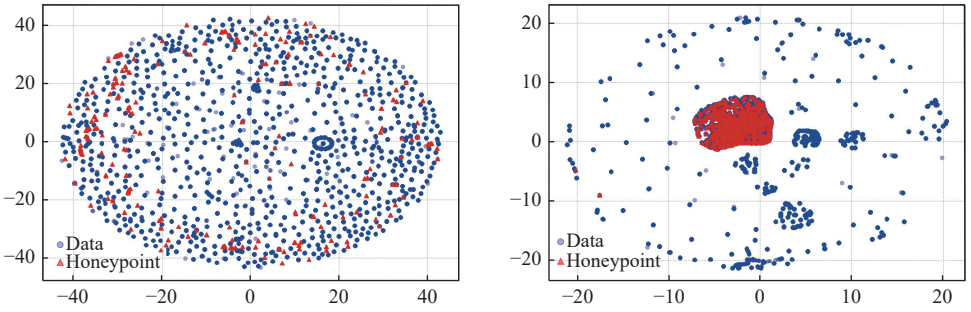


Fig. 6 Results of *PSNR* and *SSIM*

图 6 *PSNR* 和 *SSIM* 结果

据结构信息构建数据集蜜点, 很好地实现与原有数据包的耦合. 图 7(b)表明对于结构信息较为集中的 Facebook_audio3.pcapng 数据集, 本文所提出的数据集蜜点的生成方式可以准确捕获原有数据的关键结构信息, 并构建出与原有数据结构相似性很高的数据集蜜点.

从图 7 能够直观得出所构建的数据集蜜点在具有不同分布特征的结构信息的数据集上均可以实现与普通数据在特征层面的难区分, 从而说明加密数



(a) bittorrent数据集 (b) Facebook_audio3.pcapng数据集

Fig. 7 Dataset honeypoint distribution diagram based on ISCXVPN2016 dataset

图 7 基于 ISCXVPN2016 数据集的数据集蜜点分布图

据模态下数据集蜜点具备隐蔽性.

4.3 鲁棒性验证

为有效验证数据集蜜点的鲁棒性，与隐蔽性验证所使用的数据集相同，本文分别使用 CIFAR-10 数据集和 ISCXVPN2016 数据集验证数据集蜜点的在嵌入后应对攻击和破坏的能力.

4.3.1 图像数据模态

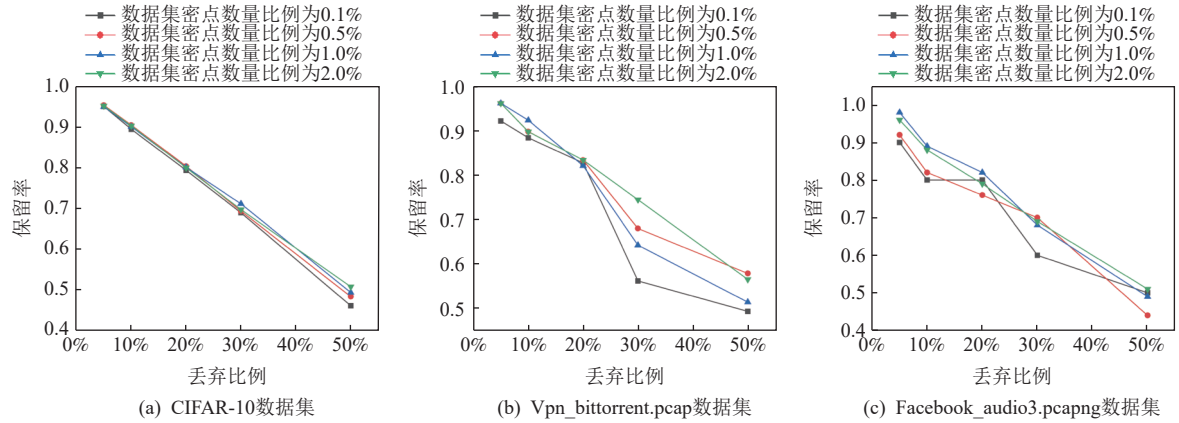
从 CIFAR-10 数据集中的每个类中随机选择 1 000 张图像，分别构造并在其中嵌入数据集总量 0.1%, 0.5%, 1% 和 2% 数据集蜜点. 依据图 2 所示的流程完成数据集蜜点的生成和嵌入.

在嵌入数据集蜜点的数据集中采用随机的方式分别丢弃总量 5%, 10%, 20%, 30% 和 50% 的数据，以模拟攻击者对图像模态的数据集蜜点的破坏. 通过计算 4.1.2 节提出的保留率衡量数据集蜜点抗破坏的能力，验证其面对攻击时的鲁棒性. 保留率计算结果如表 5 所示，可以观察到随着丢弃比例的逐渐提升，图像类数据集蜜点保留率有所降低，当丢弃比例达到 50% 时，4 种比例的数据集蜜点保留率均超过 0.44，即仍有接近一半的数据集蜜点可以与数据集紧密耦合，证明图像类数据集蜜点面对攻击展现出良好的鲁棒性. 为方便直观呈现保留率变化效果，绘制如图 8(a)所示折线图，展现图像类数据集蜜点保留率随丢弃比例的变化关系.

Table 5 Results of Robustness Verification

表 5 鲁棒性验证结果

数据集（数据量）	数据集蜜点数量（比例）	丢弃比例				
		5%	10%	20%	30%	50%
CIFAR-10（10 000）	10（0.1%）	0.900	0.800	0.800	0.600	0.500
	50（0.5%）	0.920	0.820	0.760	0.700	0.440
	100（1.0%）	0.980	0.890	0.820	0.680	0.490
	200（2.0%）	0.960	0.880	0.790	0.690	0.510
Vpn_bittorrent（422 098）	421（0.1%）	0.921	0.884	0.827	0.561	0.492
	2 109（0.5%）	0.962	0.898	0.833	0.679	0.578
	4 219（1.0%）	0.962	0.923	0.821	0.641	0.513
	8 440（2.0%）	0.961	0.897	0.833	0.744	0.564
Facebook_audio3（643 049）	642（0.1%）	0.950	0.894	0.792	0.688	0.460
	3 214（0.5%）	0.953	0.904	0.803	0.692	0.482
	6 429（1.0%）	0.950	0.901	0.801	0.711	0.492
	12 859（2.0%）	0.951	0.903	0.799	0.696	0.506



(a) CIFAR-10数据集 (b) Vpn_bittorrent.pcap数据集 (c) Facebook_audio3.pcapng数据集

Fig. 8 Retain rate of dataset honeypoint based on CIFAR-10 dataset

图 8 基于 CIFAR-10 数据集的数据集蜜点保留率

4.3.2 加密数据模态

Vpn_bittorrent.pcap 和 Facebook_audio3.pcapng 数据集分别包含超过 42 万和超过 64 万的数据包. 与在 CIFAR-10 数据集上开展的鲁棒性验证相同, 分别在 2 个数据集文件中构造占数据集总量 0.1%, 0.5%, 1% 和 2% 的数据集蜜点, 并按照空间位置生成算法实现数据集蜜点的嵌入.

在嵌入数据集蜜点的文件中分别以 5%, 10%, 20%, 30% 和 50% 的比例丢弃其中的数据, 计算数据集蜜点的保留率, 计算结果的统计值如表 5 所示. 加密模态数据集蜜点在 2 种结构信息差异较大的数据集上, 均可以实现丢弃比例达到 50% 时具备超过 0.46 的保留率, 从而展示了加密数据模态数据集蜜点所具备的鲁棒性. 依据表 5 的结果, 将加密数据模态数据集蜜点保留率随丢弃比例的变化关系绘制如图 8(b)和图 8(c)所示.

通过在 2 个数据模态下的鲁棒性实验表明, 本文所提出的数据集蜜点方法均展现出针对恶意丢弃等攻击行为的鲁棒性. 在数据丢弃比例高达 50% 时, 数据集蜜点仍能以超过 0.44 的保留率与原始数据集紧密耦合.

4.4 性能验证

本文使用 CIFAR-10 数据集衡量数据集蜜点生成阶段所产生的时间和内存开销, 以验证本文方法的性能. 随机从数据集中选取 10 000 张图片作为初始数据集. 因 PVD 算法是一种图像隐写算法, 与数据集蜜点生成阶段在逻辑上具备很高的相似性, 因此本文选取数据集蜜点生成阶段与 PVD 算法进行性能比较, 具体结果如表 6 所示.

Table 6 Results of Performance Verification

表 6 性能验证结果

算法	指标	生成比例		
		1%	5%	10%
数据集蜜点	时间/s	1.716 7	3.671 7	6.259 1
	内存/MiB	32.0	32.8	34.4
PVD	时间/s	1.614 0	4.088 3	7.078 7
	内存/MiB	34.0	34.2	34.6

通过表 6 得出, 在 3 种生成比例下, 数据集蜜点几乎均展现出相较于 PVD 算法更少的时间和内存开销, 并随着生成数据比例的逐渐升高, 数据集蜜点的性能开销相较于 PVD 算法更缓慢, 从而验证数据集蜜点具备低性能开销的特点.

5 结 论

本文提出了基于数据集蜜点的抗损毁数据标识技术, 即一种适用于数据跨域流转领域的数据标识生成和嵌入方法, 借助网络安全保障领域的“护卫”模式的思想实现数据集蜜点的生成, 并设计基于属性信息的空间位置生成算法完成数据集蜜点的嵌入. 解决数据集层面标识“易损毁”和标识“难嵌入”的问题, 并通过对数据集蜜点展开隐蔽性、鲁棒性和性能开销 3 个维度的验证实验, 证明数据集蜜点具备良好的效果. 未来工作主要包括 3 个方面: 首先, 我们期待开展更加通用化的数据集蜜点生成方式的研究, 致力于针对多模态场景下, 进一步提升数据集蜜点的生成效果. 其次, 我们希望能够强化数据集属性信息提取的能力, 实现更精准的信息获取. 最后, 我们期待将数据集蜜点应用到实际数据跨域流转过程中, 以实际测试多模态数据场景下数据集蜜点的真实表现.

作者贡献声明: 李浩波负责论文撰写并完成实验; 李默涵提出了算法思路和实验方案并修改论文; 陈鹏参与算法的设计和论文的修改; 孙彦斌对论文方法提出指导意见; 田志宏负责整体规划并审阅最终论文.

参 考 文 献

[1] Li Tao, Yang Anjia, Weng Jian, et al. Industrial Internet data disclosure audit scheme based on smart contracts[J]. Journal of Software, 2023, 34(3): 1491−1511 (in Chinese)
(李涛, 杨安家, 翁健, 等. 基于智能合约的工业互联网数据公开审计方案[J]. 软件学报, 2023, 34(3): 1491−1511)

[2] Qin Zhiguang, Xiong Hu, Wu Shikun, et al. A survey of proxy re-encryption for secure data sharing in cloud computing[J]. IEEE Transactions on Services Computing, 2016: 1−1

[3] Chen Xiao, Huang Muhong, Tian Yifan, et al. Internet of vehicles data sharing scheme via blockchain sharding[J]. Journal of Computer Research and Development, 2024, 61(9): 2246−2260 (in Chinese)
(陈骁, 黄牧鸿, 田一凡, 等. 基于分片区块链的车联网数据共享方案[J]. 计算机研究与发展, 2024, 61(9): 2246−2260)

[4] Feng Chaosheng, Liu Bin, Guo Zhen, et al. Blockchain-based cross-domain authentication for intelligent 5G-enabled Internet of drones[J]. IEEE Internet of Things Journal, 2022, 9(8): 6224−6238

[5] Tian Zhihong, Fang Binxing, Liao Qing, et al. From self-defense to protection: Suggestions for the construction and development of

network security assurance system in the new era[J]. Chinese Journal of Engineering Science, 2023, 25(6): 96−105 (in Chinese)
(田志宏, 方滨兴, 廖清, 等. 从自卫到护卫: 新时期网络安全保障体系构建与发展建议[J]. 中国工程科学, 2023, 25(6): 96−105)

[6] Wang Rui, Yang Changjiang, Deng Xiangdong, et al. Development of deception defense technology and exploration of its application in large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1230−1249 (in Chinese)
(王瑞, 阳长江, 邓向东, 等. 欺骗防御技术发展及其大语言模型应用探索[J]. 计算机研究与发展, 2024, 61(5): 1230−1249)

[7] Wang Y R, Madnick S E. A polygen model for heterogeneous database systems: The source tagging perspective[C]//Proc of 16th Int Conf on Very Large Data Bases. Australia: Proceedings, 1990: 519−538

[8] Ram S, Liu Jun. Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling[M]. Berlin: Springer, 2006: 17–29

[9] Ram S, Liu Jun. A new perspective on semantics of data provenance[C]//Proc of Int Conf on Semantic Web in Provenance Management. Aachen: CEUR-WS. org, 2009: 35−40

[10] Liu Feng, Zhang Xiaolin. Review of scientific data metadata standards and research on their universal design[J]. Modern Library and Information Technology, 2015(12): 3−12 (in Chinese)
(刘峰, 张晓林. 科学数据元数据标准述评及其通用化设计研究[J]. 现代图书情报技术, 2015(12): 3−12)

[11] Wang Fengyang, Xu Quanjun, Liu Feng, et al. Scientific data traceability description model and specification design and thinking[J]. Scientific Research Information Technology and Application, 2017, 8(1): 27−34 (in Chinese)
(王逢阳, 徐全军, 刘峰, 等. 科学数据溯源描述模型及规范设计与思考[J]. 科研信息化技术与应用, 2017, 8(1): 27−34)

[12] Zhang Qiuxia, Tian Run. Comparative study on the Internet of things identification system and ecode encoding standard[J]. China Automatic Identification Technology, 2021(1): 57−66 (in Chinese)
(张秋霞, 田润. 物联网标识体系与 Ecode 编码标准比较研究[J]. 中国自动识别技术, 2021(1): 57−66)

[13] Ren Yuzheng, Xie Renchao, Zeng Shiqin, et al. Overview of industrial Internet identifier resolution system[J]. Journal of Communications, 2019, 40(11): 138−155 (in Chinese)
(任语铮, 谢人超, 曾诗钦, 等. 工业互联网标识解析体系综述[J]. 通信学报, 2019, 40(11): 138−155)

[14] Wu Dongya. Analysis of object identifier (OID) technology and application[J]. Information Technology and Standardization, 2010(8): 66−68 (in Chinese)
(吴东亚. 对象标识符(OID)技术和应用分析[J]. 信息技术与标准化, 2010(8): 66−68)

[15] Deng Weihua, Liu Ju. Industrial Internet identification and GS1 coding system[J]. China Automatic Identification Technology, 2023(2): 40−43 (in Chinese)
(邓伟华, 刘桔. 工业互联网标识与 GS1 编码体系[J]. 中国自动识别技术, 2023(2): 40−43)

[16] Wang Baoyun. A review of Internet of things technology research[J]. Journal of Electronic Measurement and Instrumentation, 2009, 23(12): 1−7 (in Chinese)
(王保云. 物联网技术研究综述[J]. 电子测量与仪器学报, 2009, 23(12): 1−7)

[17] Luo Xiyang, Zhan Ruohan, Chang Huiwen, et al. Distortion agnostic deep watermarking[C]//Proc of 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13548−13557

[18] Yoo K, Ahn W, Jang J, et al. Robust multi-bit natural language watermarking through invariant features[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 2092−2115

[19] Liu Aiwei, Pan Leyi, Hu Xuming, et al. A semantic invariant robust watermark for large language models[J]. arXiv preprint, arXiv: 2310.06356, 2023

[20] Jia Zhaopeng, Fang Binxing, Liu Chaoge, et al. A review of network deception technology[J]. Journal of Communications, 2017, 38(12): 128−143 (in Chinese)
(贾召鹏, 方滨兴, 刘潮歌, 等. 网络欺骗技术综述[J]. 通信学报, 2017, 38(12): 128−143)

[21] Qin Xingsheng, Jiang F, Cen Mingcan, et al. Hybrid cyber defense strategies using Honey-X: A survey[J]. Computer Networks, 2023(230): 109776

[22] Yuill J, Zappe M, Denning D, et al. Honeyfiles: Deceptive files for intrusion detection[C]//Proc of the Fifth Annual IEEE SMC Information Assurance Workshop. Piscataway, NJ: IEEE, 2004: 116−122

[23] Zhang Hengru, Gong Jie. Research and design of network attack and defense platform based on virtual honeynet[C]//Proc of 2010 Int Conf on Computational and Information Sciences. Piscataway, NJ: IEEE, 2010: 507−510

[24] Bartos K, Sofka M, Franc V. Optimized invariant representation of network traffic for detecting unseen malware variants[C]//Proc of 25th USENIX Security Symp. Berkeley, CA: USENIX Association, 2016: 807–822

[25] Fu Chuanpu, Li Qi, Shen Meng, et al. Realtime robust malicious traffic detection via frequency domain analysis[C]//Proc of the 2021 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2021: 3431−3446

[26] Wu Wenling, Feng Dengguo. Research status of block cipher working modes[J]. Chinese Journal of Computers, 2006, 29(1): 21−36 (in Chinese)
(吴文玲, 冯登国. 分组密码工作模式的研究现状[J]. 计算机学报, 2006, 29(1): 21−36)

[27] Microsoft. A theorem prover from microsoft research[CP/OL]. 2022[2024-06-05]. https://github.com/Z3Prover/z3

[28] Ding Yi, Wu Guozheng, Chen Dajiang, et al. DeepEDN: A deep-learning-based image encryption and decryption network for Internet of medical things[J]. IEEE Internet of Things Journal, 2020, 8(3): 1504−1518

[29] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[R]. Toronto: University of Toronto, 2009

[30] Gerard D, Arash H, Mamun M, et al. Characterization of encrypted

and VPN traffic using time-related[C]//Proc of the 2nd Int Conf on Information Systems Security and Privacy. Portugal: SciTePress, 2016: 407–414

[31] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[32] Van der M L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 2579−2605

[33] Wu Dachun, Tsai W. A steganographic method for images by pixel-value differencing[J]. Pattern Recognition Letters, 2003, 24(9-10): 1613−1626



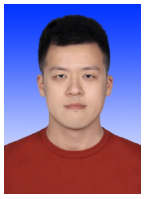
Li Haobo, born in 2000. Master candidate. Student member of CCF. His main research interests include data security and deception defense.

李浩波, 2000 年生. 硕士研究生. CCF 学生会员. 主要研究方向为数据安全、欺骗防御.



Li Mohan, born in 1987. PhD, professor. Senior member of CCF. Her main research interests include AI security, data governance, and intrusion detection.

李默涵, 1987 年生. 博士，教授. CCF 高级会员. 主要研究方向为人工智能安全、数据治理、入侵检测.



Chen Peng, born in 1988. PhD, postdoc. Member of CCF. His main research interests include explainable AI, shape manifold analysis, and data governance.

陈　鹏, 1988 年生. 博士，博士后. CCF 会员. 主要研究方向为可解释人工智能、形状流形分析、数据治理.



Sun Yanbin, born in 1987. PhD, professor. Senior member of CCF. His main research interests include network security, industrial control security, future network, and data governance.

孙彦斌, 1987 年生. 博士，教授. CCF 高级会员. 主要研究方向为网络安全、工控安全、未来网络、数据治理.



Tian Zhihong, born in 1978. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include network attack and defense confrontation, APT detection and traceability, industrial control security, and data governance.

田志宏, 1978 年生. 博士，教授，博士生导师. CCF 杰出会员. 主要研究方向为网络攻防对抗、APT 检测与溯源、工控安全、数据治理.