DataCon: 面向安全研究的多领域大规模竞赛开放数据

郑晓峰[1,2], 段海新[1,2], 陈震宇[2], 应凌云[2], 何直泽[2], 汤舒俊[2], 郑恩南[2], 刘保君[1], 陆超逸[1], 沈凯文[1], 张 甲[1], 陈 卓[2], 林子翔[2]

[1]清华大学网络科学与网络空间研究院 北京 中国 100084
[2]奇安信科技集团 北京 中国 100088

摘要 网络安全数据是开展网络安全研究、教学的重要基础资源,尤其基于实战场景下的安全数据更是科研教学成果更符合安全实践的保障。然而, 由于网络安全的技术变化快、细分领域多、数据敏感等原因, 寻找合适的网络安全数据一直是研究者们进行科研和老师开展实践教学时关注的重要问题。本文总结并分析了多个领域的经典公开安全数据集, 发现其在研究应用时存在数据旧、规模小、危害大等不足; 克服安全数据领域选择、大规模实战数据获取、安全隐私开放等困难, 构造了更符合当前科研需求 DataCon 安全数据集。数据集大规模覆盖 DNS、恶意软件、加密恶意流量、僵尸网络、网络黑产等多个领域, 且均来自实战化场景,并基于 DataCon 竞赛平台将其开放给参赛者和科研人员。目前, DataCon 数据集涵盖了已成功举办四届的 “DataCon 大数据安全分析大赛”的全部数据, 大赛被国家教育部评为优秀案例, 并进入多所高校研究生加分名单, 数据内容也一直随着真实网络环境中攻防场景的变化而持续更新。目前, DataCon 数据集涵盖了已成功举办四届的“DataCon 大数据安全分析大赛”的全部数据, 大赛被国家教育部评为优秀案例, 并进入多所高校研究生加分名单, 数据内容也一直随着真实网络环境中攻防场景的变化而持续更新。数据集持续收到科研人员、学术的数据使用申请, 支撑了多篇学术论文的发表, 充分说明了其有效性和可用性。我们希望 DataCon 数据及竞赛能够对网络安全领域产、学、研结合有所帮助和促进。

关键词 DataCon; 安全研究; 开放数据; 竞赛
中图法分类号 TP309 DOI 号 10.19363/J.cnki.cn10-1380/tn.2024.01.09

DataCon: Open Dataset for Large-scale Multiple Fields Security Research and Competitions

ZHENG Xiaofeng[1,2], DUAN Haixin[1,2], CHEN Zhenyu[2], YING Lingyun[2], HE Zhize[2], TANG Shujun[2], ZHENG Ennan[2], LIU Baojun[1], LU Chaoyi[1], SHEN Kaiwen[1], ZHANG Jia[1], CHEN Zhuo[2], LIN Zixiang[2]

[1] Institute of Network Science and Cyberspace, Tsinghua University, Beijing 100084, China
[2] QI-ANXIN Technology Group Inc, Beijing 100088, China

Abstract Cyber security data is an essential resource for cyber security research and teaching, especially the security data based on real-world scenarios is a guarantee that research and teaching results are more consistent with security practices. However, due the rapid changing technology, multiple sub-fields, data sensitivity and other reasons in the field of cyber security, finding the appropriate data has been the essential concern of researchers conducting their research and teachers teaching practice. In this paper, we summarize and analyze the classical public security data in several fields, and find that there are deficiencies in research and teaching applications such as outdated data, small data set size, and large security hazards, and overcome the difficulties of security data fields selected, large-scale real-world data acquisition, and security privacy openness, and construct DataCon security data set that is more suitable for current research needs. The DataCon data set covers DNS, malware, encrypted, malicious traffic, botnet, underground industry data set and other fields on a large scale and all from real-world scenarios, and it is open to the participants and researchers based on the DataCon competition platform. At present, the DataCon dataset covers all the data of “DataCon Big Data Security Analysis Competition”, which has been held successfully for four years, and the competition has been evaluated as an excellent case by the Ministry of Education of the People's Republic of China and has been entered into the list of extra points for graduate school of many colleges and universities, and the data content has been continuously updated along with the changes of attack and defense scenarios in the real network environment. The dataset continues to receive data use applications from scientific researchers and academics, supporting the publication of multiple academic papers, fully demonstrating its effec-

通讯作者: 陈震宇, Email: chenzhenyu@qianxin.com。
本课题得到国家自然科学基金课题资助(No. U1836213, No. U19B2034)。
收稿日期: 2021-04-26; 修改日期: 2022-11-04; 定稿日期: 2023-09-30

tiveness and usability. We hope that the DataCon data and competition can help and promote the combination of industry, academia and research in the field of cybersecurity.

Key words DataCon; security search; open dataset; competitions

1 引言

近几年来, 数据开放共享逐渐成为一种趋势, 极大促进了相关领域研究、应用的发展。然而, 网络安全行业的数据与用户的安全和隐私非常密切, 使得产业界数据开放处于封闭、滞后状态。虽然有研究人员通过主动采集、仿真生成等方式获得并公开了多个经典的安全数据集, 为安全研究的进步做出了诸多贡献, 但这些公开数据依旧难以满足当前安全研究的需求。将产业界的真实安全数据通过合理方式开放, 能够加强产、学、研结合, 有力地促进技术提升。

作为国内首个以大数据安全分析为目标的开放赛事平台, DataCon[1]克服数据获取、开放等方面的诸多挑战构建并开放了多个领域的大规模、高价值、高真实 DataCon 安全数据集, 用于支持实战化对抗场景分析比赛以及各类型的科研、教学。在 2019-2022 年成功举办四届“DataCon 大数据安全分析竞赛”并持续更新 DataCon 安全数据集, 竞赛和数据支撑产生了多个有价值的工作。

本文其余章节组织架构如下: 第 2 节介绍了现有安全数据集状况; 第 3 节总结了构建 DataCon 安全数据集面临的挑战及其包含的五个应用领域安全数据状况; 第 4 节分别对 DataCon 竞赛平台及数据集的安全开放保障进行了说明; 第 5 节基于实际的开放赛事对数据集的分析使用情况进行说明; 最后第 6 节总结全文工作并展望下一步规划。

2 现有安全数据集现状

数据集是进行科学研究的重要资源, 其质量对研究成果有着重要的影响。本节将对各个安全领域中较为经典的公开安全数据集进行介绍, 并说明其支持当前科研的不足之处。

2.1 经典公开安全数据集

KDD CUP 99 数据集[2]及其衍生数据集 NSL-KDD[3]被广泛应用于入侵检测领域科研论文的相关实验[4-11]。该数据集是 1999 年 KDD CUP 的竞赛数据, 基于 DRARPA 98 数据[12](即美国国防部高级研究规划署在麻省理工学院林肯实验室实施入侵检测评估项目生成的高仿真 TCPdump 网络连接和系统审计数据)进行一定的加工和预处理后获得, 以“连接”为基本记录单位。“连接”是在一个固定的时间间隔内, 源 IP 到目标从开始到结束的 TCP 数据包。数据集的时间跨度为 9 周, 其中 7 周约 500 万条记录作为训练数据和 2 周约 200 万条记录作为测试数据。每条“连接”记录都有 41 个固定的特征属性; 此外, 训练数据有 1 种正常的标识类型 normal 和 22 种攻击类型(如表 1 所示), 测试数据则包含更具有现实性未知的攻击类型。

表 1 KDD CUP 99 数据集

Table 1 KDD CUP 99 dataset

攻击类型	含义	详细标识
Normal	正常记录	Normal
DOS	拒绝服务攻击	Back, land, neptune, pod,smurf, teardrop
Probing	监视和其他探测活动	Ipsweep, nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write, guess, passwd, imap, multihop, phf, spy、warezclient、warezmaster

theZoo[13]是在 GitHub 获得 6200 多个 star 的恶意软件分析开源项目。该项目由 Yuval tisf Nativ 于 2014 年 1 月创建, 目前由 Shahak Shalev 进行维护, 旨在通过安全可访问的形式提供各个版本恶意软件的开放分析使用。目前为止, 该项目包含 237 个二进制形式的恶意软件样本, 80 个疑似原始恶意软件源代码, 6 个可逆向的恶意软件源代码。每个恶意软件目录包含四个文件: 加密 ZIP 存档的恶意软件文件、加密恶意软件的 SHA256 编码、加密恶意软件的 MD5 编码和存档密码。除此之外, DAS MALWERK[14]提供了 Robert Svensson 从互联网收集的 600 多个可执行恶意软件; Contagio[15]是 Mil 收集、公开的 30 多个各类型可执行恶意软件样本。

CTU-13-数据集[16]是 2011 年捷克 CTU(Czech Technical University in Prague)大学在 MCFP(The Malware Capture Facility Project)中捕获的网络流量数据, 包括僵尸网络流量、正常流量、背景流量。该数据集包含 13 个不同僵尸网络样本的捕获, 每种情况都是通过长期执行一种特定的恶意软件并在执行期间持续监测、采集相关流量数据。每一类僵尸网络数据原始流量都存储在对应的 pcap 文件, 预处理后的所有流量数据(包括标签和 argus 生成的双向 netflow 文件)存储在 biargus 文件。

Alex Top 100 万域名数据[17]和开放恶意域名数

据(如奇安信威胁情报中心 IOC 域名[18]、ZeusDGA[19]等)常常被用于可疑域名检测分析、入侵检测、web 应用防护等领域的研究, 以开放恶意域名作为黑样本, Alexa Top 除恶意域名外的域名作为白样本。Alex Top 100 万域名数据有近 100 万条数据记录, 每条记录包括域名及其在某时间的静态排名; 奇安信威胁情报中心 IOC 域名包含已证实 APT 恶意域名及其所属组织等信息。

UCI 机器学习库提供了两个钓鱼网站数据集: Mohammad 数据集[20]和 Abdelhamid 数据集[21]。Mohammad 数据集在 2015 年发布, 包含有 2456 个钓鱼网站实例, 每个实例有 30 个不同属性, 目前已被访问 15.1 万次。Abdelhamid 数据集在 2016 年发布, 通过不同来源收集了 1353 个网站数据, 其中包含 548 个合法网站、702 个钓鱼 URL 和 103 个可疑 URL, 目前已被访问 8.6 万次。此外, SofaSofa 钓鱼欺诈网站识别数据集[22]公开了通过爬虫获取的 10086 个网站训练样本和 7000 个预测样本, 每条样本记录都包含有 18 个特征变量。

2.2 现有数据集的不足

上述经典公开安全数据集对于网络安全研究起到了很大的促进和帮助作用, 然而, 它们已经难以满足现在更实时、全面、无危害的安全研究需求。本节将从数据陈旧、有效规模小、危害大三个方面对此论证。

2.2.1 数据陈旧

前述大多数据集的生产与采集时间距今已有多年。近几十年来, 网络技术飞速发展, 如果数据集的产生时间较早则根本无法有效反映当前的网络安全状况, 进而导致基于此的安全研究成果与实际情况偏差较大。

章节 2.1 中共提到 13 个数据集, 其中 8 个为单次发布, 5 个为持续更新。8 个单次发布数据集的已产生时间分布如图 1 所示, 25%的数据集产生时间在 20 年前, 50%的数据集已产生 5～10 年, 仅有 1 个产生 1 年的数据集还是时效性极高的钓鱼网站数据(生存周期通常以天为单位)。显然, 这些较为老旧的数据已无法准确地反映对应领域的当前状况。

2.2.2 有效规模小

现有数据集的有效规模较小, 主要包括数据集的整体规模小和数据集的有效数据少两大类。如果分析的数据规模远远小于实战场景下的海量数据规模, 则很可能得出较为片面的结果和认知。

数据规模小, 是指该数据集规模远远小于该类型数据的体量。例如, 2 个 UCI 钓鱼网站数据集量级均为千、SofaSofa 钓鱼欺诈网站数据量级为万, 远远小于每日新增的 50 万钓鱼网站数量[23]; theZoo、DAS MALWERK、Contagio 等恶意软件样本的数据量级更是仅在数十到数百之间。

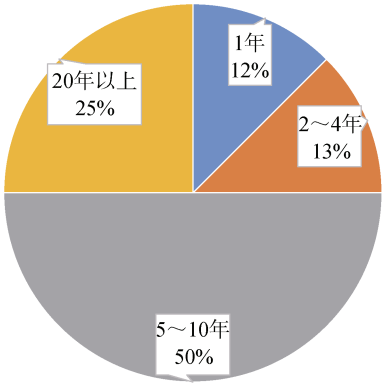


图 1 现有数据集的产生时间

Figure 1 The generate time of exist data set

数据集的有效数据少, 是指数据集中包含大量的冗余数据, 有效信息只占有较低的比例。例如 KDD99 数据集和 CTU-13-数据集中的原始流量中包含大量的冗余数据, 甚至有人针对冗余数据进行分析、优化和提炼[3,7]。

2.2.3 危害性大

危害性主要存在于恶意软件样本数据。现有的公开恶意软件样本数据通常包含可执行的恶意软件源代码、文件, 其中不乏已造成重大损失和危害的恶意软件。虽然发布者在进行开放共享时通常会通过免责声明、使用说明的方式建议使用者只将恶意软件用于研究并在运行时将其限制在未联网环境或虚拟机。但是, 恶意软件被使用者获取后的实际用途并不受数据发布者控制。免责申明、使用说明只能代表数据发布者的想法, 不能限制好奇使用者将获取的恶意软件样本非正常使用, 对自身或它人造成危害、损失; 甚至, 恶意使用者可以将恶意软件开放渠道作为工具中心获取大量的攻击工具来提升其实施各类型恶意行为的能力。

3 DataCon 安全数据集

针对安全研究、竞赛分析的实际需求和现有数据集在使用中的不足, DataCon 安全竞赛平台紧紧围绕帮助培养大数据安全人才的目标, 克服各种挑战, 构建了更符合当前需求场景的高实战、多领域大规模覆盖的 DataCon 安全数据集。

3.1 安全数据集构建面临的挑战

DataCon 安全数据集构建主要面临着选择获取

数据和数据安全开放两个方面的挑战。

3.1.1 数据的选择和获取

选择哪些细分领域的安全数据来支持安全研究和竞赛分析是需要面对的第一个问题。针对该问题, DataCon 委员会整合奇安信科技集团和清华大学资源, 协调多个业务、研究部门的业务技术专家, 从防御者、研究者的视角精挑细选确定 DataCon 安全数据集的五个领域, 同时支撑竞赛平台大数据安全分析比赛的五个赛道。五个领域分别是: DNS 数据、恶意软件数据、加密恶意流量数据、僵尸网络数据、网络黑产数据。

确定数据集领域后要解决的第二个问题是如何大规模获取这些细分安全领域的高实战原始数据。经过协调多个业务、研究部门的实战场景资源, 在 DNS 方向每年获取真实 DNS 请求信息; 在恶意软件方向, 持续捕获现网恶意软件; 加密恶意流量数据方向, 投入奇安信技术研究院天穹沙箱运行每年最新采集的恶意和正常软件并持续采集筛选其产生的流量; 僵尸网络方向, 同样通过部署公网开放蜜罐进行数据采集; 网络黑产方向, 投入奇安信采集的恶意网站域名、链接与正常网站域名、链接数据。通过上述各方向的持续投入和采集, 获得 DataCon 数据集的原始数据。

3.1.2 数据开放使用的挑战

原始安全数据直接开放使用会面临多方面的风险, 如用户隐私泄露风险、恶意软件传播风险等。原始安全数据往往包含用户或第三方的各类型行为记录隐私和身份标识信息, 将其直接开放, 不仅带来用户隐私信息泄露的风险, 而且违反《个人信息保护法》等相关的法律法规。恶意软件样本通常具备一定破坏性, 不做限制直接开放传播, 既可能被使用者恶意使用或不当使用危害它人计算机信息系统, 又可能违反《刑法》、《治安管理处罚法》、《计算机病毒防治管理办法》等相关法律法规中关于涉嫌破坏计算机信息系统的内容。

通过隐私保护等措施降低风险的同时如何确保数据可用性更提升了数据开放的隐私保护难度。大数据分析与原始数据特征息息相关, 一旦脱敏过程中破坏数据特征将会严重影响大数据分析的结果, 从而无法满足数据开放支撑研究和竞赛分析的初衷。因此, 在对前述情况进行脱敏处理时, 需要尽可能保护数据可用性, 不影响数据集的研究、分析使用效果。详细脱敏工作介绍见第 4.2 章内容。

3.2 数据集状况

接下来将以 2020 年数据集为例对 5 个细分领域的 DataCon 数据进行详细介绍。

3.2.1 DNS 数据

DataCon DNS 数据集[24]是经过处理的部分 2020 年 3～5 月真实 DNS 请求信息, 包括三个不同的子数据集。DNS 数据集 1 是来自 1000 个恶意域名以及约 20000 个请求量与之相似的干扰域名的 DNS 请求信息, 信息内容具体包含客户端 IP 信息、域名、解析结果、相关域名 whois 信息等。DNS 数据集 2 分为训练集和测试集, 训练集包含约 2000 个有标签黑白域名的 DNS 请求信息, 测试集包含 10000 多个无标签黑白域名的 DNS 请求信息, 信息内容具体包含客户端 IP 信息、域名、解析结果等。DNS 数据集 3 是来自约 10000 个无标签黑白域名的 DNS 请求信息, 信息内容具体包含客户端 IP 信息、域名、解析结果、TTL 等。不同数据子集各自存放在一个文件目录下, 目录中包含的文件名、内容解释、数据量如表 2 所示, 各文件的字段介绍如表 3 所示。

3.2.2 恶意软件数据

DataCon 恶意软件数据集[25]源自每天从现网捕获的恶意代码, 分为训练集和测试集两部分, 训练集中包含 6000 个有标记的恶意软件样本文件(2000 个黑样本为明确的挖矿型恶意代码, 4000 个白样本为明确的非挖矿型恶意代码), 测试集包含 6000 个未

表 2 DataCon_DNS 数据内容及规模概况

Table 2 Overview of content and size for DataCon DNS data

文件名	解释	DNS 数据集 1	DNS 数据集 2		DNS 数据集 3
			训练集	测试集	
access.csv	客户端访问域名记录, 按小时聚合	2291340	3705372	3705372	9721367
flint.csv	域名解析记录, 包括被解析到的域名(cname ...)的 A 和 AAAA 记录	1987608	2203261	2203261	2203261
fqdn.csv	所有待分类域名和 flint rdata 中出现的域名	20513	3839	17469	17469
ip.csv	所有客户端 IP 和 Flint 解析结果中出现的 IPv4 的信息	419156	44607	1871018	1871018
ipv6.csv	所有客户端 IP 和 Flint 解析结果中出现的 IPv6 的信息	11976	4671	35638	35638
label.csv	数据标签	477	1989	—	—
whois.json	相关域名 whois 信息	162664	—	—	—

表 3 DataCon_DNS 数据字段解释

Table 3 The field of DataCon DNS data

文件名	字段	字段解释	涉及 DNS 数据集
access.csv	fqdn_no	域名编号	1、2、3
	encoded_ip	加密后的 IP	1、2、3
	count/ request_cnt	请求量	1、2、3
	time	时间	1
	date	日期	2、3
	hour	小时	2、3
	total_request	该时间段全网请求量	2、3
flint.csv	fqdn_no_x	域名编号	1、2、3
	flintType	解析类型	1、2、3
	encoded_value	解析结果(域名编号或加密 IP)	1、2、3
	requestCnt	访问量	1、2、3
	date	日期	1、2、3
fqdn.csv	TTL	生存时间值	3
	fqdn_no	域名编号	1、2、3
	encoded_fqdn	域名字符串特征	1、2、3
ip.csv/ ipv6.csv	encoded_ip	加密 IP	1、2、3
	country,	IP 地理信息	1、2、3
	subdivision,	IP 地理信息	1、2、3
	city	IP 地理信息	1、2、3
	latitude,	IP 经纬度	1、2、3
	longitude	IP 经纬度	1、2、3
	isp	ISP 信息	1、2、3
label.csv:	family_no	恶意域名家族	1
	label	黑白域名标签	2

标记的待检测恶意代码样本文件。单个恶意代码样本的大小主要在 20KB 至 10MB 之间, 样本的总大小约为 12GB。

为确保样本多样性, 基于百万个样本的恶意样本集进行相似性分析, 过滤掉相似样本后, 最终获得全部 12000 个样本。

此外, 为避免样本运行, 样本 PE 结构中的 MZ 头、PE 头、导入导出表等区域均已抹去, 虽然无法动态分析, 但其代码指令特征依然存在。

3.2.3 加密恶意流量数据

DataCon 加密恶意流量数据集[26]源自于 2020 年 2 月～6 月收集的恶意软件与正常软件, 经奇安信技术研究院天穹沙箱运行并采集其产生的流量筛选生成。本数据集定义的恶意流量为恶意软件(均为 exe 类型)产生的加密流量, 白流量为正常软件(均为 exe 类型)产生的加密流量。流量内容为 443 端口产生的 TLS/SSL 数据包。

数据集包括训练集和测试集, 训练集规模为 3000 个有标注的 pcap 文件(其中黑样本、白样本数量均为 1500), 测试集规模为 2000 个待检测 pcap 文件, 每个 pcap 文件都是一个恶意软件在一个客户端 IP 产生的流量数据, 不同 pacp 文件代表不同恶意软件产生的恶意流量。训练集和测试集的黑样本分别为 2020 年 2 月～2020 年 5 月和 2020 年 6 月捕获的恶意软件加密流量, 所有白样本均为 2020 年捕获的正常软件加密流量。

3.2.4 僵尸网络数据

DataCon 僵尸网络数据集[28]包括两个子数据集, 分别是相同获取来源僵尸网络样本文件数据集和 HTTP 蜜罐数据集。

僵尸网络样本文件数据集, 来自部分公网开放蜜罐捕获到的僵尸网络文件投递行为。分两个部分, 第一部分为僵尸网络文件样本, 第二部分为样本间的来源关系, 即被同一个 IP 投递且下发地址为同一个 IP, 这样的样本大概率会有代码上的相似特性, 部分可以确定为同一套源码编译。僵尸网络样本文件数据均为 32 位 ELF。样本文件“mips”为 MIPS 大端可执行文件, 样本文件“x86”为 Intel IA-32 小端可执行文件。每个类型各 512 个文件, 且对于每个文件而言, 都有另一个文件集中的一个文件和它“同源”(即

从同一套源码编译而来或在同一套源码之上略作改动编译而来)。

HTTP 蜜罐数据集来自部分公网开放蜜罐捕获的 HTTP 请求流量, 数据采集的时间范围是 2020 年 5 月。存储形式为一个 2.53GB 的 honeypot.json 文件, 包含 2745694 行记录, 文每一行为一个 JSON 字符串, 代表一条蜜罐日志, 即针对蜜罐的一次 HTTP 请求, 详细字段解释如表 4 所示:

表 4 蜜罐数据字段介绍

Table 4 The field of honeypot data

字段	解释
eventid	事件类型, 数据文件中均为“attack.web”
method	HTTP method
body	HTTP body
url	HTTP request URL
header	HTTP request headers
path	HTTP request path
protocol	应用层协议, 数据文件中均为“http”
src_ip	源 IP 地址(已加盐哈希处理)

3.2.5 网络黑产数据

DataCon 网络黑产数据集[28]源自于奇安信在 2020 年 7 月下旬～8 月上旬期间收集的恶意网站域名、链接与正常网站域名、链接数据。本数据集共分为 12 个数据子集, 每个数据子集包含一个.txt 格式的域名、URL 混合列表文件以及一个 csv 格式的域名备案信息文件。域名备案信息包括 8 个特征项: 域名、网站备案号、位名称、单位性质、审核时间、网站名称、网站地址、详细地址。本数据集及各子集的数据规模详情如所示。

表 5 DataCon 网络黑产数据集状况

Table 5 The scale of underground industry dataset

数据(子)集	域名\URL 记录数	域名备案记录数
1	214408	5324
2	219760	5439
3	211585	5279
4	204471	4824
5	195540	5331
6	168535	4506
7	155058	4560
8	268709	6331
9	207358	5608
10	191588	3776
11	219221	5116
12	478543	8679
全部	2734776	64773

由于网络黑产数据是完全实战化场景, 数据集中网站网络情况以及运营情况变化无法预测, 导致同一域名或同一 URL 在不同时间所展示的内容不一样, 所以本数据集不提供具体的黑产分类信息。本数据集涉及的黑产类别包含且不限于如下类别: 涉赌、涉黄、涉毒、涉枪、涉诈、传销、接码平台、账号买卖、个人信息买卖、黑客相关、发卡平台、空包、卡池猫池、网赚、游戏私服、流量劫持、政府仿冒、假证买卖、支付平台、跑分平台、IDC 服务商、CDN 服务商等, 其中还包含大量的正规网站劫持事件。部分网络黑产类别定义如表 6 所示。

表 6 部分网络黑产类别定义

Table 6 The definition of part underground industry

黑产类别	含义
账号买卖	指的是网络虚拟账号买卖、租售, 包括且不限于游戏账号、微信支付宝账号等
个人信息买卖	指买卖公民个人身份信息, 包含四件套(银行卡、手机号、身份证、U 盾)、隐私信息(包含切不限于快递信息、人脸信息等)、手持身份证照片等
黑客相关	指买卖黑客工具、黑客接单等
支付平台	指三方支付、四方支付、聚合支付等
涉诈	包含且不限于各类杀猪盘、股票配资、理财、贷款诈骗、假冒正规交易、贷款网站(如咸鱼、转转、度小满)等等

3.3 数据集优势

DataCon 数据集应用于安全研究和比赛分析场景具有多方面的优势, 如高实战性、多领域大规模覆盖、脱敏开放(低危害性)等。

3.3.1 高实战性

DataCon 数据集来源于实战、服务于实战, 为相关研究和分析提供了应用场景和实施方式的案例支撑。其高实战性, 主要通过三个方面来体现: 来源真实、数据新鲜、热点业务。

来源真实, 即数据集的所有原始数据都是从现网的实战业务环境中采集和获取, 而不是通过构建仿真系统生成。仿真系统通常只能考虑到真实环境中的一部分影响要素, 生成仿真数据和现网抓取数据相比存在一定偏差。基于仿真数据进行研究和分析, 其发现成果必然会和实际状况有所偏差, 实战场景应用效果同样会受到影响。直接使用源自于现网的真实数据可以有效的避免这一问题, 提升研究和分析效果。

数据新鲜, 即数据集来自于最近时间段(每年竞赛中都会开放最新数据), 可以较好地反映相关领域的当前状况。随着技术不断地更新、升级和迭代, 各

个领域产生的数据及其特征同样会随之变化。采集时间较久的数据, 其数据特征与当前实际状况存在偏差, 研究和分析结果同样会受到影响从而与实际状况产生偏差。基于较为新鲜的数据进行研究和分析, 能够更好地反映该领域的当前状况。

热点业务, 数据集的数据采集领域都是较为热点的领域, 受到广大安全从业者和攻击者关注。基于热点业务的数据集进行研究, 能在更大的范围内影响当前网络空间安全态势。

3.3.2 多领域大规模覆盖

DataCon 数据集涉及领域范围全面、数据有效规模大, 实现了安全数据的多领域大规模覆盖。

如章节 3.2 所述, DataCon 数据集数据涉及多个安全领域, 囊括了 DNS 数据、恶意软件数据、加密恶意流量数据、僵尸网络数据、网络黑产数据五个领域方向的数据。DNS 是互联网基础协议之一, 一直是互联网通信的重要研究内容, 以此数据为基础能够从主流防御者的角度进行分析考察。恶意软件同样是非常传统的安全领域, 各类木马、病毒、勒索软件、挖矿软件等感染了越来越多的互联网用户并造成大量危害, 相关领域的研究分析能帮助促进恶意软件的快速检测发现。加密恶意流量数据则是加密通信前提下对恶意软件数据进行分析检测, TLS 等部署成本越来越低, 越来越多的恶意软件使用加密作为主要传输手段, 相关流量监测分析成为恶意软件发现的新战场。僵尸网络一直是进行 DDoS 等网络攻击事件的基础, DDoS 防御工作会长期与僵尸网络的研究和监控紧密相关。网络黑产受巨大利益驱使, 不仅搭建各类黑灰产网站, 还为了引流、提升搜索引擎排名攻击正规网站, 对黑产网站进行分析研究, 才能够达成知己知彼知威胁并实施有效的黑产发现、打击。

上述各个领域数据子集的规模在对应领域公开数据集中都处于领先地位, 足以支撑相关的大数据安全分析。DNS 数据集, 涉及约 4 万个域名的 3200 多万条各类型记录信息, 现有经典公开数据集中从未出现过恶意软件数据包含 6000 个恶意软件样本, 要远远超过现有公开数据集中的数百个。加密恶意流量数据包含 5000 个有标注的 pcap 文件。僵尸网络数据则包含 2745694 行 HTTP 蜜罐数据以及 1024 个存在同源关系的僵尸网络样本。网络黑产数据包括 273 万域名 URL 信息和 6.4 万域名备案信息。

3.3.3 脱敏开放

DataCon 数据集在开放使用之前会根据领域实际情况对各个数据子集进行相关的脱敏操作。典型脱敏情况如下所示: 数据包含产生该数据的用户标识符信息, 则对其身份信息进行隐私保护处理; 数据包含大量的第三方标识符信息, 同样需要对其进行隐私保护处理; 软件存在危害性, 则需要脱敏破坏其可执行性。在对上述情况进行脱敏处理时, 还需要尽可能保护数据可用性, 不影响数据集整体的研究、分析使用。详细脱敏工作介绍见第 4.2 章内容。

4 竞赛平台及数据的安全开放

DataCon 安全竞赛平台, 是国内首个以大数据安全分析为目标的开放赛事平台, 为多种安全分析竞赛提供平台支撑、相关数据的安全开放及安全交流社区生态。

4.1 竞赛平台

为确保竞赛公平公正, 竞赛平台在注册报名、赛题发布、答案提交等基础功能外, 还提供了具备弹性可扩展和数据保护能力的虚拟化执行环境。

竞赛过程中, 平台将根据赛事方向确定虚拟化环境配置及测试代码流程和功能的样例数据, 并根据各方向参与人数同方向所有参赛选手各自提供相同配置的虚拟化环境。选手可在运行环境中自行配置第三方库等依赖环境、调试竞赛代码, 并在持久化目录下存放环境配置脚本、竞赛程序代码, 以免因为运行环境重置导致数据丢失。

竞赛算法运行检测时, 为避免解题方案执行受到外部因素影响或恶意代码对外部环境造成影响, 虚拟环境会断开网络连接。评委根据代码执行结果及过程进行最终评判, 能够对选手的解题思路和方法进行充分考察。

4.2 数据安全开放

依托 DataCon 开放竞赛平台, 各项竞赛能向学术界提供真实的脱敏数据资源, 开放协作, 在以产学研深度融合推进安全领域的实战性研究成果转化, 为网络空间安全的发展创造更大价值。

4.2.1 身份标识符隐私保护

DataCon 安全数据集中典型的身份标识符/准标识符信息主要是 IP 地址、MAC 地址、域名信息等。源 IP 地址、MAC 地址能够对产生信息的用户身份信息进行唯一标识, 进而将其他数据字段及其中隐含的隐私信息与用户身份关联泄露具体用户的隐私。目标 IP 地址、域名信息则能够将各项信息与确切的信息目标流向相互关联, 从而泄露目标 IP、域名的相关内容隐私。针对上述问题, DataCon 数据集结合数据与应用场景相关状况, 采用多种措施对身份标识字段进行脱敏处理, 从而实现身份标识符隐私保护。

加密恶意流量数据中, 客户端 IP 的主要作用是对加密恶意流量产生源进行标识, 因此对其处理方式是将原始 IP 映射为内网网段 IP, 即完全保留了加密恶意流量产生源的标识区分能力, 又保护了原有的加密恶意流量产生源 IP 信息。

DNS 数据中, 域名即具备一定的特征信息, 又是访问流量、解析日志中不同记录的目标身份标识。对其进行隐私保护处理时, 充分考虑到了域名自身的特征信息用于进一步分析的可能, 将每个域名映射为一个域名代码和相应的字符串特征说明, 充分保留了原始域名的字母、数字、词语、特殊符号、顶级域名、长度等信息。隐私保护处理规则如下: 顶级域名和特殊符号保持不变, a 表示字母、0 表示数字、[aaa]表示 aaa 为一个词语。如: 原始域名 abchello-12.com 隐私保护后的特征码为 aaa[aaaaa]-00.com。此外, DNS 数据中 IP 信息(客户端 IP 和域名解析 IP 地址)既具备特征信息, 又是区分不同流量来源、解析目标的标识符, 此外还是关联 IPwhois 信息库的唯一标识。为此, 在对其进隐私保护处理时, 将前 3 段映射为加密字符串仅保留第 4 段, 并提供加密 IP 的国家、省、市、经纬度、运营商等信息。在不泄露用户隐私(IP 访问记录)的基础上保留了 IP 的唯一标识能力和第四段分布特征, 并满足了关联 IP whois 的基本查询需求。

此外, 根据数据领域的实际情况, 还进行 IP 加密、选择特定端口流量等多样性的数据隐私保护措施。

4.2.2 软件脱敏处理

恶意软件或僵尸网络样本文件的原始文件通常具备危害性和可执行性, 不经过处理即公开发布, 可能被恶意使用者当作恶意危害他人的攻击工具, 或被好奇的使用者使用无意中造成自身或他人的损失。针对上述问题, DataCon 数据集针对不同的样本数据和应用场景, 对相关软件样本进行脱敏处理, 从而实现软件危害性脱敏。

恶意软件领域会提供大量可移植可执行文件(Portable Executable, PE)文件样本以供研究分析和比赛使用。PE 文件是目前 Windows 平台上的主流可执行文件格式, 包括可执行程序 EXE 文件、动态链接库 DLL 文件等, 将其安全开放需要破坏可执行性并保留研究价值。分析 PE 文件格式可知, MS-DOS 头、PE 头、导入导出表等区域会涉及到样本的运行。例如: MS-DOS 头包含 MZ 头信息、PE 头偏移地址等信息且能够调用 PE 头; PE 头包含 PE 文件标识、标准头、扩展头等信息; 导入表(IMAGE_DIRECTORY_ENTRY_IMPORT)会提供 PE 文件加载时依赖的 DLL 及填充所需函数的地址; 导入地址表(IMAGE_DIRECTORY_ENTRY_IAT)则是填充真正的函数地址。清洗上述区域信息后, PE 样本文件虽然无法运行进行动态分析, 但其恶意行为的指令特征依然存在, 不影响各项静态分析的效果。

僵尸网络方向用以研究分析和比赛使用的开放数据是 ELF 文件样本。ELF 文件是 Linux 平台上的主流可执行文件格式, 其文件段(Section)信息一般包括代码段(.text)、只读数据段(.rodata)、已初始化全局数据段(.data)、未初始化全局数据段(.bss)、符号表(.symtab)等。结合僵尸网络样本的分析考察内容, 去除了样本中的只读数据段等内容, 保留了代码段信息, 处理后数据依旧不影响不同僵尸网络样本同源分析等场景分析使用。

5 基于开放赛事的数据集分析

基于 DataCon 竞赛平台和安全数据集, 奇安信集团、清华大学和蚂蚁集团联合主办了国内首个以大数据安全分析为目标的大型比赛“DataCon 大数据安全分析竞赛”。当前已成功举办四届(2019—2022)比赛, 接来下将以 2020 年赛事为例, 从比赛结果、分析方法机理两个方面来对数据的有效性、价值进行说明。

5.1 赛事情况及结果分析

DataCon2020 大数据安全分析竞赛基于安全数据集的细分数据领域, 设置了五大安全赛道来解决不同场景下的安全问题。具体赛题包括: DNS 恶意域名分析和恶意代码分析、加密恶意流量检测、网络黑产分析以及僵尸网络分析等重要的攻防实战场景。各个赛道的题目和规则不尽相同, 但都旨在选手能够自由发挥、充分扩展解题思路。此外, 还通过埋入特殊采分点、代码检查等多种反作弊方式来确保比赛公平性。

DNS 域名方向包括 6 道赛题, 最终得分由各赛题得分以不同权重累加计算。如图 2 上图展示了 40 多支提交答案参赛战队的赛题得分状况色度图, 战队与赛题对应网格颜色越深代表该战队在该赛题得分越高; 下图展示了总成绩得分最高的 20 只队伍得分情况, 在多个得分区间存在激烈的竞争。成绩分布在不同色度(区间)说明赛题难度和评分具备一定合理性; 不同队伍能取得相近成绩(相近色度、分数区间)说明赛题数据能够有效支撑不同解题方法。

恶意软件方向采用淘汰制, 如图 3 所示, 资格赛共 37 名队伍提交有效答案, 30 支队伍获得晋级资格;

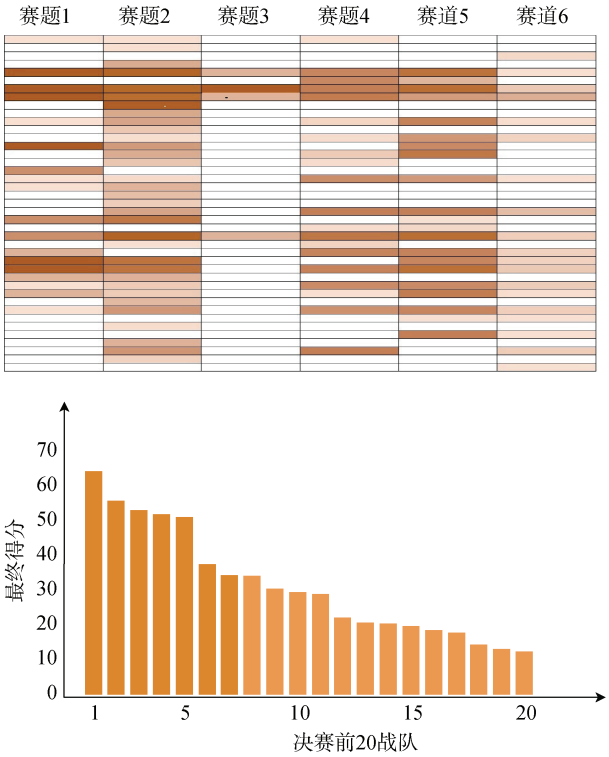


图 2　DNS 域名方向-各赛题不同战队得分色度图及决赛排名前 20 战队最终得分

Figure 2　DNS domain direction-the score chromaticity diagram of each question and the final score of top 20

经过初赛和附加赛的筛选，前 14 支队伍进入决赛；决赛中多支队伍使用各自方法成功解决问题。

该方向各个阶段都有多支队伍获得较好成绩，同样说明该方向数据集能够在安全条件下有效支持不同的分析方法。

加密恶意流量检测方向同样采取淘汰制，参赛队伍资格赛及决赛得分状况如图 4 所示，左图资格赛中多支晋级队伍的成绩聚集在 100 分、85 分、70 分三个分数段；右图决赛中的各队伍得分主要分布在 60～85 区间。不同队伍的分数分布状况说明该方向数据即使经过处理同样能够支持多个采分点的评估设置及不同解题方法的应用。

僵尸网络方向是由背景知识题和不同难度的三道赛题(赛题 3 包括主观分和客观分)组成。图 5 左图展示了参赛队伍的各题得分状况色度图，共有 32 支队伍成功提交答案，由背景知识题得分状况可知既有相关背景的队伍也有无背景知识的队伍，虽然有背景知识的队伍通常成绩更好，但也有一些无背景知识的队伍同样取得较好的成绩，这也说明了该方向数据的隐私保护处理并没有受到背景知识的局限，能够支持新的分析思路和方法。

网络黑产方向题目为完全实战化场景，目标网站网络情况以及运营情况变化无法预测，导致同一

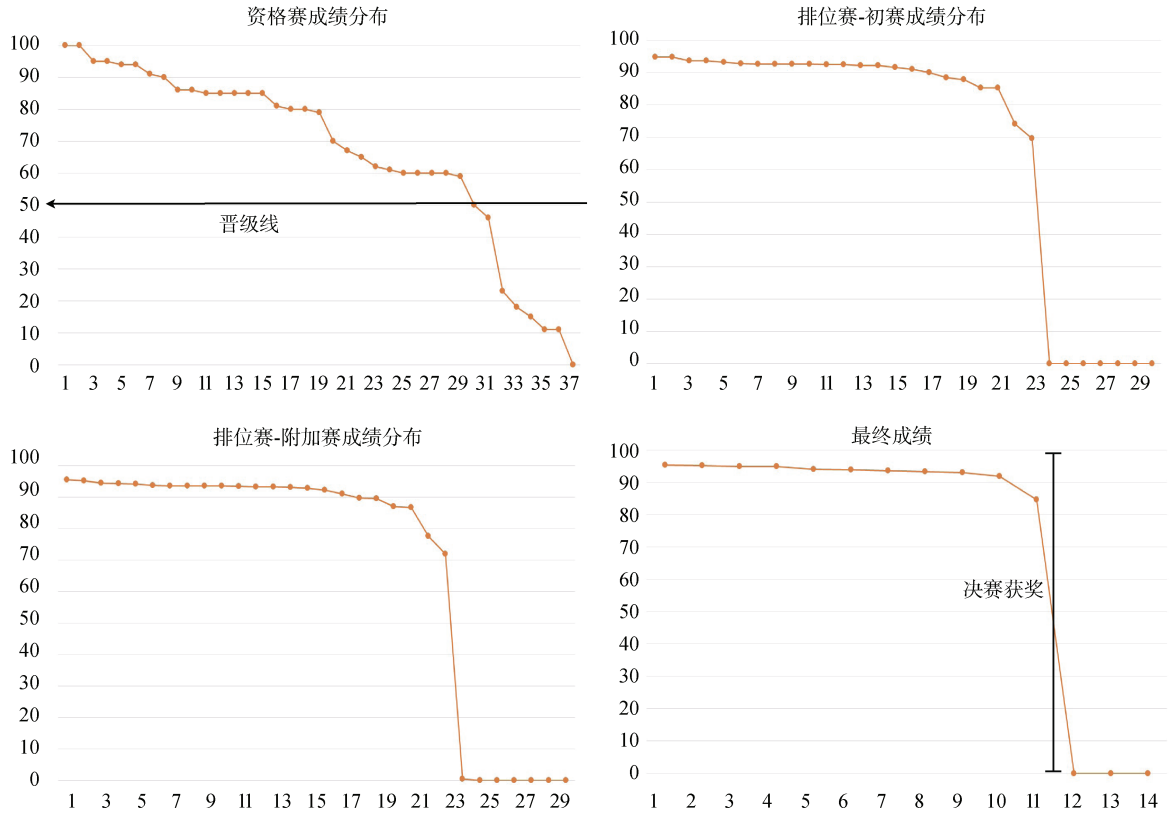


图 3　恶意软件方向-各阶段成绩状况(纵轴为分数，横轴为战队成绩排名)

Figure 3　Malware direction-the results of each stage (the vertical axis is the score, the horizontal axis is ranking)

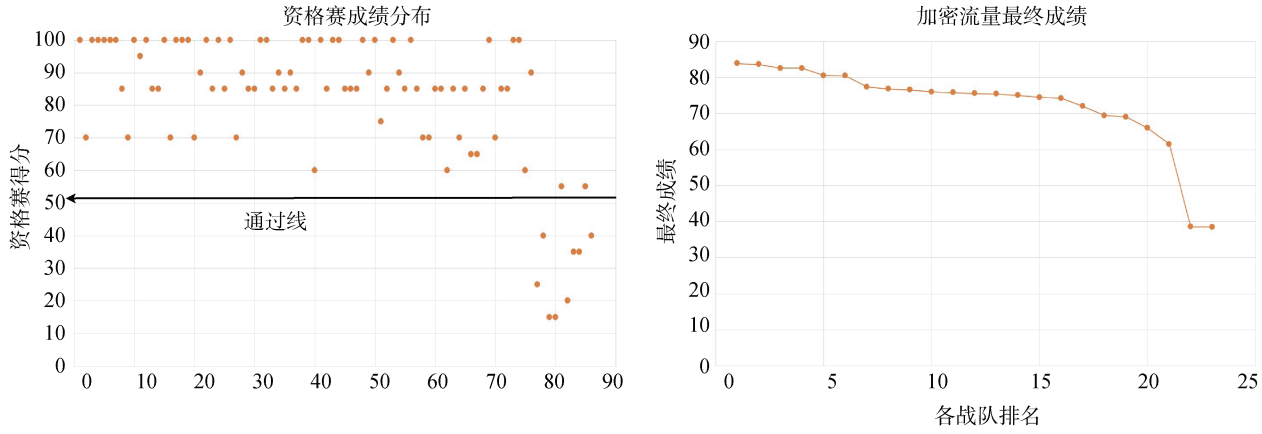


图 4　加密恶意流量方向-资格赛和最终成绩分布图

Figure 4　Encrypted malicious traffic- Direction-distribution of qualification and final score

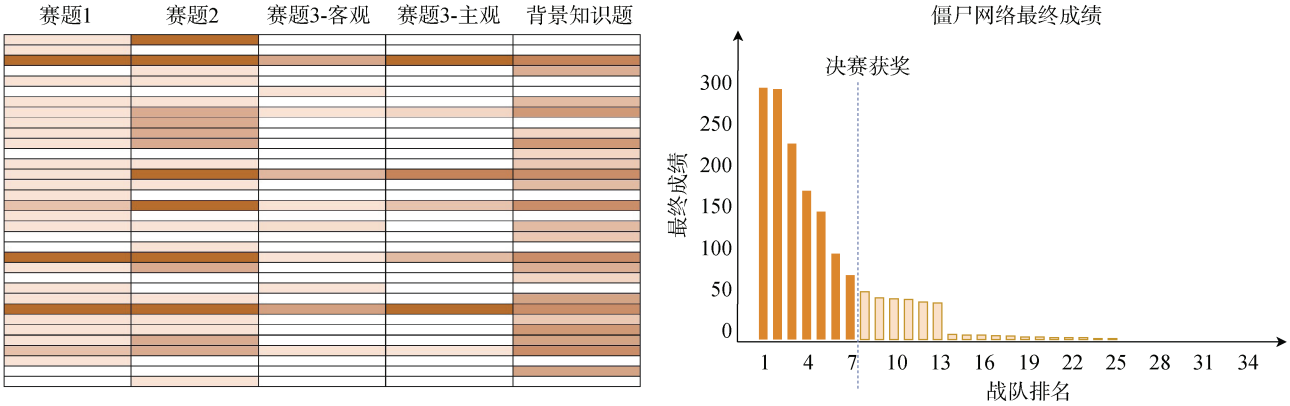


图 5　僵尸网络方向-各题得分色度图及决赛得分状况

Figure 5　Botnet direction-the score chromaticity diagram of each question and the final score

域名或同一 URL 不同选手在不同时间所展示的内容不一样, 所以该题由评委老师结合选手的分类结果及 writeup 进行综合评分, 最终得分状况如图 6 所示, 不同队伍都产出了有效的黑产分析结果, 同样论证了数据的真实价值。

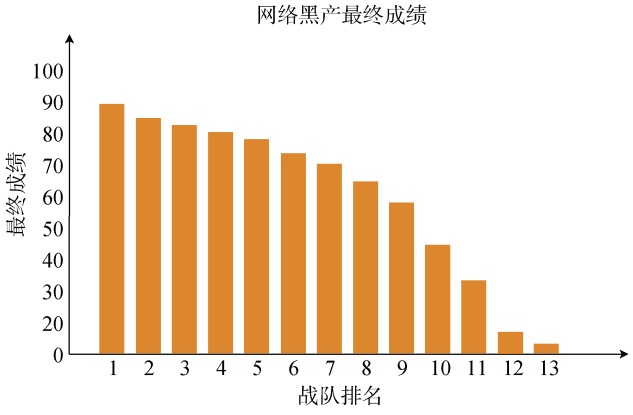


图 6　网络黑产方向最终成绩

Figure 6　Underground industry direction- the final result

5.2　优秀分析方法及数据有效性

竞赛的各个赛道方向均产出了优秀的分析实践, 对 DataCon 平台及安全数据集在安全研究方面的实战价值提供了有力的结果支撑。因篇幅有限, 此处仅结合部分优秀分析实践对数据有效性进行说明介绍。

DNS 方向, 结合广州大学 IStar 战队分析恶意域名家族的解题思路[29]对数据有效性进行说明。该解题过程为将问题定位为类别严重不平衡的多分类问题题进行数据预处理、特征工程和建模分析, 过程中涉及特征包括域名字符串特征、域名解析 IP 数量、域名解析 IP 分布、域名解析 IP 变化频度等。DNS 脱敏数据特征如章节 4.2.1 所述, 域名脱敏保留了原始域名的字母、数字、词语、特殊符号、顶级域名、长度等信息; IP 脱敏保留了唯一标识能力和第四段特征、IPwhois 信息。比对分析可知脱敏数据直接满足分析过程所需的 3 项数据特征, 仅影响 IP 分布特征, 但 IP 分布特征可通过附加的 IP whois 运营商信息部分替代。综上所述, DNS 数据集在域名分析领域

具有接近原始数据的分析价值且不会泄露域名 IP 的隐私关系。

恶意软件分析方向，结合中科院信工所 IIE- AntiMiner 战队的分析思路[30]对数据有效性进行说明。分析思路为：首先根据黑样本(挖矿软件)的行为特性初步确定待关注特征，然后根据不同特征采用灰度图、直方图、静态特征模型等分别进行处理和验证，最后基于不同模型效果和特点建立更稳定的融合模型(图 7)。分析过程中主要使用 PE 文件的二进制字节、调试信息、重定位信息、PE 头基本信息、导出表个数和名称、Section 名称大小属性等、字符串特征，进而提取匹配路径、注册表、URL、IP 地址、比特币钱包地址、挖矿软件常见字符串等特征。前述软件样本脱敏通过 MZ、PE、导入导出表破坏可执行性，但 PE\0\0 的基本静态特征可恢复，仅缺导入表信息。总而言之，本方向的脱敏数据不支持恶意软件样本的动态分析，但支持静态分析方法且对分析结果影响很小。

加密恶意流量检测方向，结合清华大学 HawkEye 战队的分析思路[31]对数据有效性进行说明。检测方法的总体结构是利用不同的异构特征训练多个分类器并通过多数投票(Majority Voting)的方式来获取最终的判定结果，整体模型如图 8 所示。分析过程中只有 IP 地址分类器会涉及脱敏的 IP 数据，但只考虑 IP 与特定 IP 的通信关系。该数据集中的 IP 脱敏方式为将其映射到内网网段，确保了 IP 的唯一标识性，完全满足分析需求。总而言之，该场景下数据基本为原始数据，仅对 IP 进行了脱敏以确保隐私，同时不影响 IP 作为标识和关联符的场景需求。

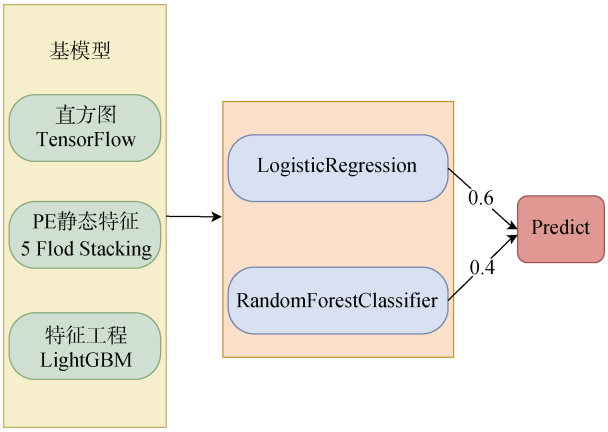


图 7 恶意软件方向 writeup-融合模型

Figure 7 Malware direction writeup-fusion model

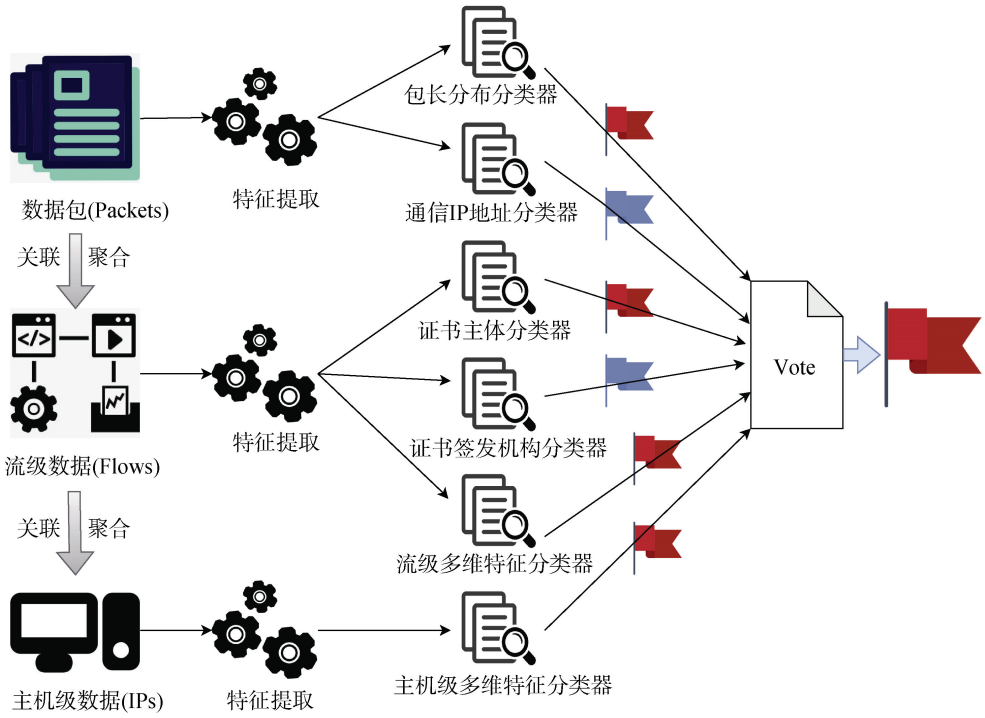


图 8 加密恶意流量检测方向 writeup-检测模型

Figure 8 Encrypted malicious traffic writeup-detection model

僵尸网络方向，结合阿里云安全战队第一题的分析验证思路[32]对数据有效性进行说明。分析思路为：首先分析 Botnet 常规模型、传播模型、触发 DNS 流量方式、攻击流量来源；然后从蜜罐流量的 RCE payload 挖掘出攻击者 IP，从样本逆向分析找出僵尸网络通信域名；最后通过域名访问拓扑关联发现其

它域名。其中寻找攻击者 IP 和僵尸网络通信域名主要基于样本代码段, 前述脱敏方式对代码段信息并无破坏和不影响使用。该分析方法最终找出全部正确结果也充分说明脱敏对该题目分析无明显影响。

网络黑产方向, 结合浙江大学 matrix 战队的黑产网站团伙发现的分析验证思路[33]对数据价值进行说明。分析思路为: 以域名作为图节点, 以 IP 地址、https 证书、域名解析记录、相同统计链接等 4 类关系作为边, 构建图模型进行关联分析及黑产判定(如图 9 所示)。该题目数据仅包含黑产域名列表和域名 ICP 信息, 分析思路中使用的其它节点关系均为选手基于域名信息通过合法方式自行扩展获得, 不涉及各种隐私。可以说, 本数据集为网络黑产相关研究提供大量的基础研究素材。

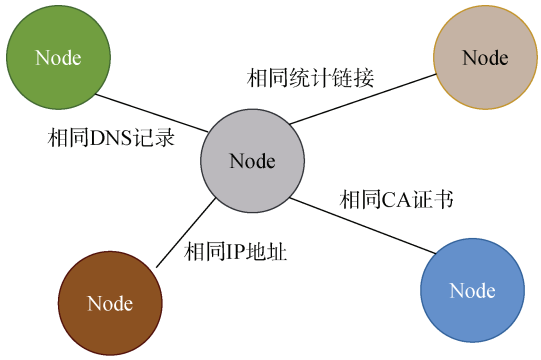


图 9　网络黑产方向 writeup-黑产网站关联模型

Figure 9　Underground industry direction writeup-the website association model

6　结论

随着大数据时代的到来, 开放数据对安全研究的价值和意义也逐渐为人们所认可。为促进安全领域研究, 我们构建 DataCon 平台及安全数据集, 并通过大数据分析竞赛形式进行了相关的推广和验证。在此过程中, 克服和解决了构建高质量安全数据集和设置低背景门槛的开放式题目两个方面的问题和难点, 构建了兼顾实战性、覆盖面、规模、安全性的数据集, 提出了既满足业务需求又满足前沿研究的考察问题。比赛参与者结合安全、大数据分析等多个角度的智慧对数据进行分析解决问题, 提出很多优秀的分析思路并取得一定成果。

接下来, DataCon 竞赛平台将积极推动从安全竞赛、数据开放、技术交流等多个维度共同构建安全数据开放、研究、交流的生态圈。这些工作的意义不仅在于夯实安全研究的基础, 更重要的是通过数据作为纽带促进产学研结合, 培养积极防御型安全人才、促进研究行业前沿研究、让网络更安全。

致　谢　感谢 2019-2022 四届 DataCon 平台主办单位及各位专家的大力支持, 感谢各位参赛选手积极参与并分享比赛思路。

参考文献

[1] DataCon. https://datacon.qianxin.com/.

[2] KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[3] The NSL-KDD Data Set. https://web.archive.org/web/20150205070216/http://nsl.cs.unb.ca/NSL-KDD/.

[4] McHugh J. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory[J]. *ACM Transactions on Information and System Security*, 3(4): 262-294.

[5] Dokas P, Ertoz L, Kumar V, et al. Data mining for network intrusion detection[C]. *Proc. NSF Workshop on Next Generation Data Mining*. 2002: 21-30.

[6] Panda M, Patra M R. Network intrusion detection using naive bayes[J]. *International journal of computer science and network security*, 2007, 7(12): 258-263.

[7] Tavallaee M, Bagheri E, Lu W, et al. A Detailed Analysis of the KDD CUP 99 Data Set[C]. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009: 1-6.

[8] Moustafa N, Slay J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)[C]. *2015 Military Communications and Information Systems Conference*, 2015: 1-6.

[9] Janarthanan T, Zargari S. Feature Selection in UNSW-NB15 and KDDCUP'99 Datasets[C]. *2017 IEEE 26th International Symposium on Industrial Electronics*, 2017: 1881-1886.

[10] Kwon D, Natarajan K, Suh S C, et al. An Empirical Study on Network Anomaly Detection Using Convolutional Neural Networks[C]. *2018 IEEE 38th International Conference on Distributed Computing Systems*, 2018: 1595-1598.

[11] Kumar V, Sinha D, Das A K, et al. An Integrated Rule Based Intrusion Detection System: Analysis on UNSW-NB15 Data Set and the Real Time Online Dataset[J]. *Cluster Computing*, 2020, 23(2): 1397-1418.

[12] 1998 DARPA INTRUSION DETECTION EVALUATION DATASET. http://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset.

[13] theZoo - A Live Malware Repository. https://github.com/ytisf/theZoo.

[14] DAS MALWERK. https://dasmalwerk.eu/.

[15] CONTAGIO. http://contagiodump.blogspot.com/.

[16] MCFP-The Malware capture Facility Project. https://mcfp.weebly.com/mcfp-dataset.html.

[17] Alexa Top 100w domains. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip.

[18] 奇安信威胁情报中心-IOC 域名. https://ti.qianxin.com/apt/

[19] zeusDGA. http://www.secrepo.com/misc/zeus_dga_domains.txt.zip.

[20] UCI Machine Learning Repository Phishing Websites Data Set. https://archive.ics.uci.edu/ml/datasets/phishing+websites#.

[21] UCI Machine Learning Repository Websites Phishing Data Set https://archive.ics.uci.edu/ml/datasets/Website+Phishing.

[22] SofaSofa-Phishing fraud website identification. http://sofasofa.io/competition.php?id=10.
(SofaSofa 钓鱼欺诈网站识别. http://sofasofa.io/competition.php?id=10.)

[23] Mobile phone security status report in 2020 Q3. http://zt.360.cn/1101061855.php?dtid=1101061451&did=610689546.
(2020 年第三季度中国手机安全状况报告, http://zt.360.cn/1101061855.php?dtid=1101061451&did=610689546.)

[24] DataCon-opendata-DNS Malicious domain. https://datacon.qianxin.com/opendata/dns.
(DataCon 开放数据-DNS 恶意域名. https://datacon.qianxin.com/opendata/dns.)

[25] DataCon-opendata-maliciouscode. https://datacon.qianxin.com/opendata/maliciouscode.
(DataCon 开放数据-恶意软件数据. https://datacon.qianxin.com/opendata/maliciouscode.)

[26] DataCo-opendata-maliciousstream. https://datacon.qianxin.com/opendata/maliciousstream.
(DataCon 开放数据-加密恶意流量数据. https://datacon.qianxin.com/opendata/maliciousstream.)

[27] DataCon-opendata-botnet. https://datacon.qianxin.com/opendata/botnet.
(DataCon 开放数据-僵尸网络数据. https://datacon.qianxin.com/opendata/botnet.)

[28] DataCon-opendata-backsite. https://datacon.qianxin.com/opendata/blacksite.
(DataCon 开放数据-网络黑产数据. https://datacon.qianxin.com/opendata/blacksite.)

[29] DataCon2020_Excellent Writeup of DNS. https://datacon.qianxin.com/blog/archives/211.
(DataCon2020_DNS 方向优秀解题思路 Writeup. https://datacon.qianxin.com/blog/archives/211.)

[30] DataCon202_Excellent Writeup of maliciouscode. https://datacon.qianxin.com/blog/archives/141.
(DataCon2020_恶意软件方向优秀分析成果 Writeup. https://datacon.qianxin.com/blog/archives/141.)

[31] DataCon202_Excellent Writeup of maliciousstream. https:// datacon.qianxin.com/blog/archives/122.
(DataCon2020 加密恶意流量方向优秀分析成果 Writeup. https://datacon.qianxin.com/blog/archives/122.)

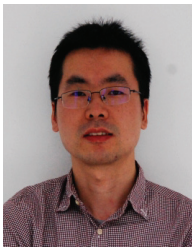
[32] DataCon2020_Excellent Writeup of botnet. https://datacon.qianxin.com/blog/archives/109.
(DataCon2020_僵尸网络方向优秀分析成果 Writeup. https://datacon.qianxin.com/blog/archives/109.)

[33] DataCon2020_Excellent Writeup of backsite. https://datacon.qianxin.com/blog/archives/196.
(DataCon2020_网络黑产方向优秀分析成果 Writeup. https://datacon.qianxin.com/blog/archives/196.)

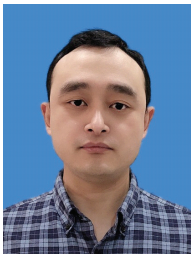
郑晓峰 正高级工程师, CCF 会员, 现任奇安信网络空间安全中心主任, 国家卓越工程师计划导师, 奇安信网络攻防领域带头人。研究领域为网络基础设施、基础协议安全、漏洞挖掘。Email: zhengxiaofeng@qianxin.com

段海新 于 2000 年在清华大学获得博士学位。现任清华大学网络科学与网络空间教授, 清华大学-奇安信联合研究中心负责人。重点关注 DNS、CDN、PKI、Web、TLS 等基础设施和基础协议的安全。Email: duanhx@tsinghua.edu.cn

陈震宇 于 2019 年在中国科学院软件研究所获得博士学位。现任奇安信技术研究院/清华大学-奇安信联合研究中心研究员。研究领域为网络基础设施测量与安全、网络空间隐私保护。Email: chenzhenyu@qianxin.com



应凌云 于 2010 年在中国科学院软件研究所获得博士学位。现任奇安信技术研究院/清华大学-奇安信联合研究中心研究员, 星图实验室负责人。研究领域为: 动态分析沙箱、软件空间测绘、软件自动攻防等。Email: yinglingyun@qianxing.com



何直泽 现任奇安信技术研究院/清华大学-奇安信联合研究中心研究员。研究领域渗透测试与漏洞利用、大网视角下的攻防对抗与威胁捕获。Email: hezhizeg@qianxing.com

汤舒俊 于 2019 年在清华大学材料科学与工程专业获得硕士学位。现任奇安信技术研究院/清华大学-奇安信联合研究中心研究员。研究领域为网络基础设施、基础协议安全、网络黑产生命周期研究、僵尸网络发现与攻防研究等。Email: tangshujun@qianxin.com

郑恩南 于 2020 年在国际关系学院通信与信息系统专业获得工学硕士学位。现任奇安信技术研究院/清华大学-奇安信联合研究中心研究员。研究兴趣包括: 加密流量分析, 攻击流量检测, 恶意软件网络行为分析等。Email: zhengennan@qianxin.com

刘保君 于 2019 年在清华大学获得博士学位。现任清华大学网络科学与网络空间研究院博士后, 博士毕业于清华大学计算机科学与技术系。研究方向为网络测量与网络安全。Email: lbj@ Tsinghua.edu.cn

陆超逸 于 2017 年在北京邮电大学信息安全专业获得学士学位。现在清华大学网络空间安全专业攻读博士学位。研究领域为网络安全、互联网测量。Email: lcy17@mails.tsinghua.edu.cn

沈凯文 于 2022 年在清华大学获得博士学位, 蓝莲花、Tea Deliverers 核心成员, 云起无垠创始人。研究领域为邮件安全、攻防渗透、协议安全。Email: skw17@mails.tsinghua.edu.cn

张甲 于 2010 年在清华大学获得博士学位。现任清华大学网络科学与网络空间研究院助理研究员。 研究领域为网络异常行为智能检测、网络基础设施安全测量、网络协议安全分析等。Email: zhangjia2017@tsinghua.edu.cn

陈卓 于 2018 年在北京邮电大学密码学专业获得硕士学位, 现任北京奇安信 A-TEAM 安全研究员, 研究领域为攻击者画像以及黑灰产相关数据分析。Email: chenzhuo0618@gmail.com

林子翔 现任奇安信涉网犯罪研究中心任研究员, 涉网犯罪研究中心互联网情报中心负责人。研究领域为网络黑灰产识别、网络攻击溯源。Email: lingzixiang@qianxin.com