arXiv:2601.05813v1 [cs.DB] 9 Jan 2026

描述：多区域云蜜网数据集（MURHCAD）*

预印本

Enrique Feito-Casares[1], **Ismael Gómez-Talal**[†1]，以及**José-Luis Rojo-Álvarez**[1,2,3]

[1]胡安·卡洛斯国王大学信号理论、通信与远程信息处理系统及计算系，西班牙马德里丰拉拉达[2]D!lemmaLab Ldt，西班牙马德里丰拉拉达 [3]胡安·卡洛斯国王大学数据、复杂网络与网络安全科学研究中心，西班牙马德里。

摘要

本文介绍一个全面、高分辨率的蜜网数据集，旨在支持对全球网络攻击行为进行独立分析。该数据集于2025年6月9日至11日在 Microsoft Azure 上连续72小时采集，包含由部署在四台地理位置分散的虚拟机上的三个蜜罐（Cowrie、Dionaea 和 SentryPeer）捕获的132,425起攻击事件。每条事件记录均包含丰富的元数据（UTC 时间戳、源/目的 IP 地址、自治系统和组织映射、地理定位坐标、目标端口及蜜罐标识符），以及派生的时间特征和标准化协议分类。我们为希望利用该数据集开展异常检测、协议滥用研究、威胁情报和防御策略设计的研究人员提供切实可行的指导。描述性统计显示出明显的偏斜：2,438个唯一源 IP 地址分布于95个国家，但排名前1%的 IP 地址占全部事件的15%；此外，三种协议占据主导地位：会话初始协议（SIP）、Telnet 和服务器消息块（SMB）。时间分析发现，协调世界时07:00和23:00出现显著的高峰时段峰值，其间穿插着维护导致的间隙，揭示了运行盲点。地理空间映射进一步凸显了平台特定的偏差：SentryPeer 捕获到集中于北美和东南亚的 SIP 洪泛攻击，Cowrie 记录的 Telnet/SSH 扫描主要来自西欧和美国，而 Dionaea 则记录了欧洲节点周边的 SMB 漏洞利用攻击。通过将精细的时间分辨率与丰富的上下文地理定位和协议元数据相结合，这一独立数据集旨在助力开展可复现、云规模的调查，研究不断演变的网络威胁。我们还提供配套的分析代码和数据访问详情，以便研究人员立即使用，并推动网络安全研究领域进一步的方法创新。

关键词 异常检测 · Azure 云部署 · 网络攻击数据集 · 蜜网 · 时间攻击模式

背景

有效网络安全解决方案的设计与评估高度依赖于真实且多样化的数据集。传统数据集往往在范围或真实性方面存在局限，通常缺少实时攻击痕迹或上下文元数据。近期的研究成果旨在通过在受控但真实的环境中提供经过整理的流量轨迹、日志和行为模式来弥补这一空白。例如，有研究提出了一个用于捕获针对易受攻击的虚拟机的全面攻击场景的数据集，其中包括 FTP 和 SSH 漏洞利用、ARP 欺骗以及暴力破解尝试，所有这些攻击均通过 Wireshark 和 tcpdump 等工具进行观测 [Shandilya 等人， 2023] 。信息物理系统也通过带有标注的数据集进行了研究，这些数据集记录了流体控制子系统中的故障、异常和网络攻击，为工业环境中的安全监控提供了宝贵的洞见 [Laso 等人， 2017]。

*本文已提交至电气电子工程师学会，等待可能的发表

†ismael.gomez.talal@胡安·卡洛斯国王大学.es

动态且大规模的项目提供了在持续数日的网络防御演习期间获取的真实网络流量和系统日志 [Tovarˇnák 等人, 2020], , 使研究人员能够在模拟的攻击—防御场景下测试异常检测方法。在传输控制协议层面，研究人员对信息物理系统和物联网（IoT）环境中的缓冲延迟进行了刻画，从而提供了有关协议层交互如何降低实时系统响应能力的细粒度数据 [Al-Hammouri 和 Olsen, 2021]。在物联网领域，设备异构性和分布式拒绝服务（DDoS）威胁带来了独特挑战，近期的数据集着重于在异构网络中对多种攻击类型进行分类和检测，例如 IoT-DH [Saif 等人, 2024]。光无线传感器网络的大规模工业数据也已开放获取，从而能够利用延迟、吞吐量和分组错误率等指标对信息物理系统的性能进行分析 [Faheem 和 Butt, 2022]。与我们的工作更为相关的是基于蜜罐的项目。其中，Hornet 40 [Valeros 和 García, 2022] 以及 Hornet 65 Niner [Valeros 和 García, 2025] 强调了地理位置在攻击暴露中的作用，汇总了来自全球部署的被动传感器和蜜罐的数百万条流记录（例如纽约、伦敦和班加罗尔），并提供标准化的 Zeek 或 Argus 日志。

为弥补这一空白，我们提出 MURHCAD，这是一个通过部署在 Microsoft Azure 上的全球分布式蜜网，在 72 小时内采集的高分辨率入站网络攻击数据集。该数据集有助于开展有关攻击突发、昼夜节律、协议级滥用，以及源 IP 地址和组织所实施恶意行为的集中情况等新颖研究。除了时间戳、协议和端口信息等标准属性外，该数据集还为每个事件补充了源地理位置、组织归属（通过 IP 地址到自治系统编号（ASN）的映射获得）以及目标主机。与以往的蜜罐数据集不同，该数据集特别强调时间粒度，使研究人员能够按小时调查攻击行为模式，并支持对协议滥用、攻击突发和时空攻击动态的分析。

尽管以往的蜜罐数据集为了解全球攻击面提供了宝贵见解，但它们往往缺乏高分辨率的时间标注、有关攻击者来源的详细元数据，以及异构平台之间的同步数据。尤其是，现有文献仍缺少一种数据集，能够将大规模、地理分布广泛的蜜罐轨迹与小时级分辨率及丰富的上下文元数据相结合，从而支持细粒度的时空分析[Ring 等人, 2019]。
所呈现的数据集 [Feito Casares 等人 2025]通过整合部署在四台地理位置分散的虚拟机上的三个不同平台（Cowrie、Dionaea 和 SentryPeer[3]）的蜜罐日志，全面描绘了现实世界中的网络攻击。蜜罐是一种通过模拟存在漏洞的系统来吸引和研究网络攻击的安全机制。该数据集细粒度的时间分辨率使研究人员能够按小时剖析攻击动态，揭示具有特征性的每日高峰时段以及维护导致的盲区。同时，协议级注释（例如会话初始协议（SIP）、Telnet 或服务器消息块（SMB））凸显了攻击者在 VoIP、传统远程登录服务和 Windows 文件共享端点之间不断转移的关注重点，为网络分段和防火墙策略设计提供了可操作的洞见。

除了时间和协议分析之外，多区域云蜜网数据集还提供了丰富的地理定位元数据，支持跨区域威胁情报分析：来自 95 个国家、超过 2,400 个不重复源 IP 地址的数据揭示了全球扫描模式，并将关注点集中于高连接性枢纽。通过将每次入侵映射到其自治系统组织和来源国，可以评估异常检测算法区分良性背景噪声与有组织扫描活动的有效性，并进一步优化适用于多分类时间序列场景的机器学习分类器。

由于该语料库在一致的三天窗口内同时涵盖定性维度（攻击类型、协议、源国家）和定量维度（时间戳、端口、纬度/经度），因此特别适合在受控且贴近现实的条件下对防御机制进行基准测试。工业从业者可利用这些见解优化蜜罐部署和自适应监控，学术研究人员则可利用这些数据开发并验证生成模型、预测工具和自动分割策略，以预判和缓解新兴网络威胁。

数据收集方法与设计

实验方案是在 Azure 云平台上跨不同地理区域部署四台虚拟机，以模拟真实环境。每台虚拟机都运行蜜罐服务，以监控并记录活动。这些虚拟机从6月9日到6月11日持续运行，从而实现全面的数据收集。

本文所述数据集 是从 Microsoft Azure 云服务提供商上的专用基础设施中提取的 r
并使用 T-Pot Co 社区版 [Telekom Security, 2024]。T-Pot 社区版是一款一体化、可选 y

[3]注册商标。

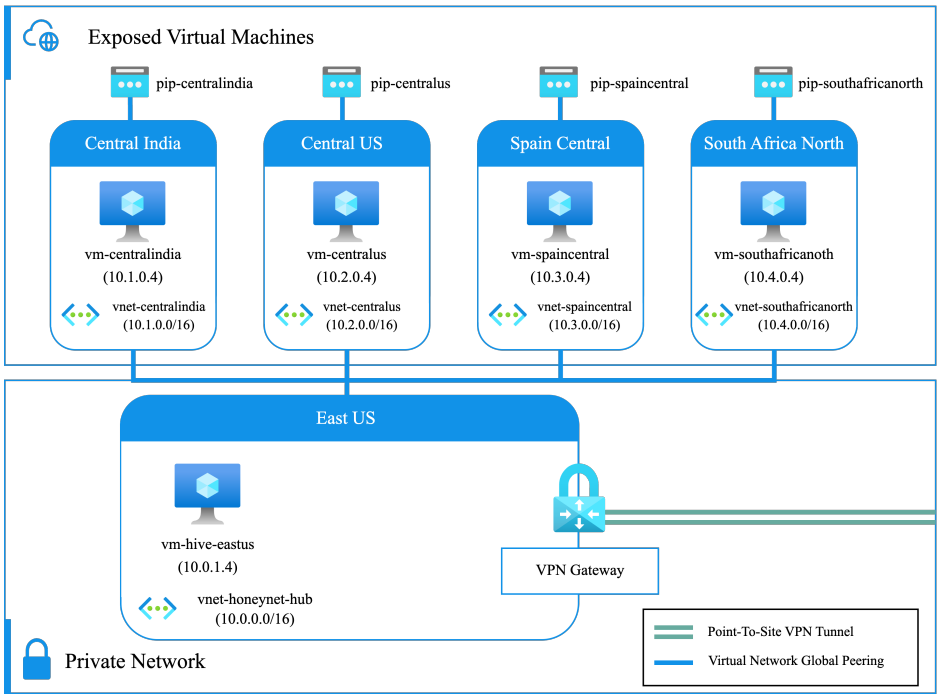


图 1：基于 Azure 的全球中心辐射式蜜网架构，包含四台暴露的传感器虚拟机，每台均位于各自的虚拟网络中，并全部与一个中央蜜网中心虚拟网络建立对等连接；该网络托管 Hive 虚拟机和 VPN 网关。

分布式、多架构蜜罐平台，支持 20 多种蜜罐，并提供丰富的可视化效果、动态实时攻击地图以及众多安全工具，以提升欺骗体验。
为捕获关键服务端口上的高价值威胁活动，我们选择并部署了三种不同的蜜罐工具，并采用镜像配置：

- Cowrie：模拟 SSH 和 Telnet 服务，提供虚拟文件系统、可自定义横幅以及交互式 Shell 响应。它支持多个并发会话，可高保真地捕获暴力破解和凭据窃取活动 [Oosterhof，2024]。
- Dionaea：模拟 SMB、MSSQL、MySQL、TFTP 和 MQTT 等服务，通过模拟网络交互捕获恶意软件和漏洞利用载荷 [DinoTools，2021]。
- SentryPeer：监控 SIP 流量中的 INVITE 泛洪和格式错误的请求，并记录源IP地址和尝试拨打的号码 [SentryPeer, 2024]。

该蜜网在 Microsoft Azure 中采用中心辐射式拓扑实现。位于印度中部、美国中部、西班牙中部和南非北部的四个传感器虚拟网络（VNet）中，部署了配置相同的可抢占虚拟机，运行 Cowrie、Dionaea 和 SentryPeer 蜜罐（见表 1）。图 1 展示了所述架构的详细示意图。每个传感器 VNet 均分配了一个 /16 地址空间，并由网络安全组保护，这些安全组仅允许在特定蜜罐服务端口上接收入站流量。每台传感器虚拟机都配置了公共 IP 接口，允许所有蜜罐服务端口不受限制地接收入站流量；而管理端口无法从互联网访问，只能通过 VNet 访问。所有传感器 VNet 均与位于美国东部的中央枢纽 VNet 建立了对等互连（见表 2）。对等互连使用 Azure 私有全球骨干网，可在不暴露于公共互联网的情况下提供低延迟连接。枢纽 VNet 包含一个配置为使用 SSL/TLS 隧道的点到站点 VPN 网关，管理员通过该网关在专用网络上与枢纽和传感器实例建立 SSH 会话。传感器实例收集的所有遥测数据均汇总到枢纽，以供摄取，从而实现集中式索引、可视化和长期归档。

除了提供的数据集外，项目代码仓库中还提供了完整的教程，介绍如何部署所述基础设施并安装所有必要工具，以复现实验 [Feito-Casares 等人， 2025]。

表 1：传感器虚拟机的配置和特征。

区域	虚拟机名称	SKU 与规格	Disk	操作系统映像	私有 IP
印度中部	vm-centralindia	标准 _D2as_v5	64 GiB	Ubuntu 24.04 LTS	10.1.0.4
美国中部	vm-centralus	标准 _D2as_v5	64 GiB	Ubuntu 24.04 LTS	10.2.0.4
西班牙中部	vm-spaincentral	标准 D2as v5_ _	64 GiB	Ubuntu 24.04 LTS	10.3.0.4
南非北部	vm-southafricanorth	标准 _D2as_v5	64 GiB	Ubuntu 24.04 LTS	10.4.0.4

表 2：Hive 虚拟机的配置和特征。

区域	虚拟机名称	SKU 与规格	Disk	操作系统映像	私有 IP
美国东部	vm-hive-eastus	标准 _B4ms	128 GiB	Ubuntu 24.04 LTS	10.0.1.4

验证与质量

对数据集进行了全面验证，以确保所收集数据的技术质量、准确性和一致性。采用了多种验证方法，以核实蜜罐基础设施和数据收集流程的完整性与可靠性。

数据集构成

多区域云蜜网数据集包含通过蜜罐系统捕获的132,425条日志，记录了2025年6月9日至6月11日期间发生的网络攻击。每一行对应一条独特的攻击交互记录，表示分布在全球不同位置的多台虚拟机上的蜜罐所进行的独立捕获。为便于理解，表3简要说明了数据集中的各个变量，包括从蜜罐网络中捕获的地理空间、时间、协议级和基础设施级指标。如表4和表5所示，该数据集包含定量变量和定性变量。

采集基础设施

在为期 72 小时的数据采集期间，基于 Azure 的基础设施始终保持所有虚拟机的高可用性。记录在案的维护时段涉及计划关机，因此在数据采集中形成了明显的间隙（在图2中表现为锯齿状模式）。

数据指标

全面的统计分析表明，各采集时段的数据质量始终如一。极端集中模式（排名前 1% 的 IP地址占事件总数的 15%）反映的是真实的攻击者行为，而非数据伪影。端口分布分析证实，临时源端口和集中于常见服务端点的目标端口呈现出真实的模式，与预期的攻击模式相符。

进一步审视分类变量（表4）可发现，多个维度的分布都高度偏斜。尽管该数据集仅涵盖三个日历日，攻击却集中发生在 6月10日（2025年6月10日），这反映了由运营因素或攻击者驱动的时间偏差。源IP地址尤为多样，共有 2,438 个不同来源；然而，排名第一的来源（23.175.48.211）单独就产生了 18,561 个事件，接近总数的 15%，这表明少数主机产生了不成比例的大量流量。自治系统分析呈现出类似模式，排名第一的自治系统编号（VDI-NETWORK）同样贡献了这 18,561 个事件。从地理分布来看，攻击源自 95 个国家，但美国占据主导地位，事件超过 38,600 起（约占 31% ）。在目标端，蜜罐仅部署在四个国家，其中印度接收到 53,057 次访问。在协议滥用方面，共出现 13 种不同协议，但会话初始协议（SIP）以 55,060 起事件居首；这些事件几乎全部由 SentryPeer 蜜罐捕获，其记录的数量也同为 55,060，凸显出 VoIP相关探测的普遍性，远超 Telnet 或 HTTP 等更传统服务。

数值摘要（表5）进一步阐明了攻击的时空动态。时段分析显示，攻击平均发生在协调世界时 11:41，但标准差高达 7.19 小时，证实了全天候近乎持续的探测活动。四分位距覆盖协调世界时 06:00–18:00，在约 07:00 和 23:00 观察到明显峰值（见图3(a)）。端口统计显示，源端口均值为 48,604，表明

表3：变量说明。

变量	说明
星期	事件发生的星期几（例如，星期二）。
date	事件的确切日期，格式为 MM/DD/YY。
源 IP 地址	攻击源所使用的公网 IP 地址。
源组织	与源 IP 地址段关联的组织（根据自治系统编号查询结果）。
源国家名称	与源 IP 关联的国家名称。
源地理 IP 地址	srcIp 的冗余值，通常用于地理定位。
目标 IP 地址	目标系统的公网 IP 地址。
目标内部 IP 地址	接收攻击的内部主机的私有 IP 地址。
目标主机名	遭受攻击的机器（虚拟机或容器）的主机名。
目标地理 IP 地址	用于确定目标地理定位的 IP 地址（与 dstIp 类似）。
目标国家名称	目标系统所在的国家（通过目的IP地址确定）。
协议	攻击期间使用的协议（例如 sip、telnet、smbd）。
攻击类型	捕获该事件的蜜罐系统：Cowrie，Dionaea，或 SentryPeer。
hour	记录事件时的小时数（0–23）。
day	从时间戳中提取的数字日期。
源端口	攻击者使用的源端口号。
srcLat, srcLon	攻击源的地理坐标（纬度和经度）。
dstPort	攻击者瞄准的目的端口。
dstLat, dstLon	目标系统的地理坐标。

表4：数据集中的类别特征摘要。其包含时段、地理和协议级属性，展示了每个特征的唯一值数量、最常见值（Top）及其频率（Freq）。

特征	唯一	Top	Freq
星期	3	星期二	76318
date	3	2025年6月10日	76318
源 IP 地址	2438	23.175.48.211	18561
源组织	522	VDI-NETWORK	18561
源国家名称	95	美国	38600
源地理 IP 地址	2438	23.175.48.211	18561
目标 IP 地址	11	135.235.171.89	47370
目标内部 IP 地址	4	10.1.0.4	53057
目标主机名	4	vm-centralindia	53057
目标地理 IP 地址	11	135.235.171.89	47370
目标国家名称	4	印度	53057
协议	13	sip	55060
攻击类型	3	SentryPeer	55060

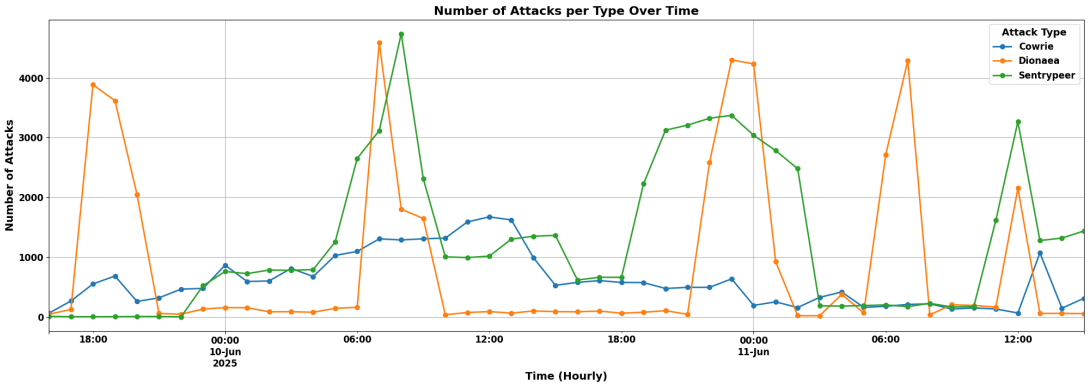
临时端口的使用 用于扫描的端口，而目标端口则呈现出极端偏斜：均值为 2,693，75t h 百分位数为 5,060（SI P)，这表明针对 VoIP 端点的攻击较为集中。

这些描述性统计凸显了当代网络空间探测的两个基本特征：攻击者来源极为多样（95 个国家的 2,400 多个 IP地址），但攻击又集中于少数顶尖来源（单个排名第一的 IP地址和自治系统编号产生了 15% 的事件，且 SIP 在协议使用情况中占主导地位）；同时，网络空间探测具有时间持续性，全天各时段均有发生，并呈现出明显的高峰时段。这些发现有助于制定自适应防御策略，既应对零星扫描形成的长尾，也应防范导致大多数入侵尝试的高频攻击者。

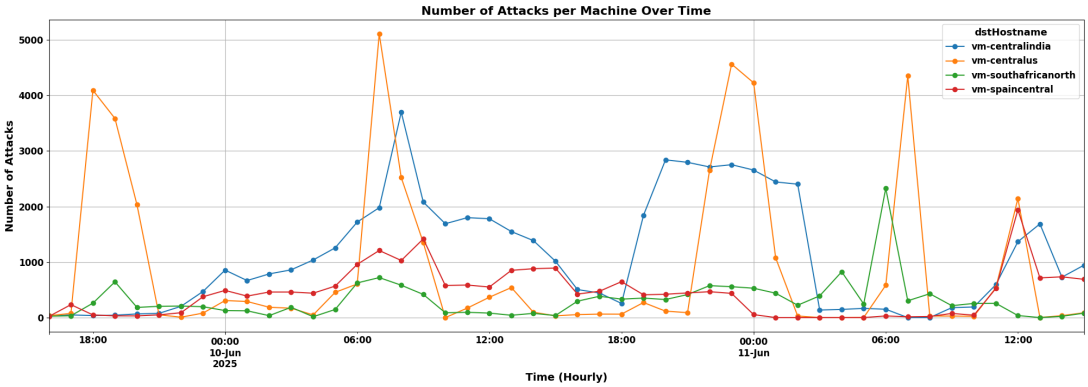
图 3(b) 展示了各蜜罐平台遭滥用的协议分布。每种协议对应的三根不同柱形分别显示了 SentryPeer、Cowrie 和 Dionaea 蜜罐如何捕获不同服务上的流量：(1) SentryPeer（蓝色柱形）记录到的攻击绝大多数针对 SIP，事件数超过 55,000，这反映出该平台专注于 VoIP相关探测和利用尝试；(2) Cowrie（橙色柱形）的流量以 Telnet 为主（约 29,000 个事件），这与其通常用作 SSH/Telnet蜜罐、模拟传统远程登录服务的用途相符；(3) Dionaea（绿色柱形）则呈现出更多样的模式，可见大量 SIP 和 Telnet 探测，不过其流量最高的目标是 SMBD (Microsoft-DS/CIFS)，约有 36,000 个事件

表5：数据集中数值特征的汇总统计信息。其中包括时间指标（例如小时和日）、源端口号和目的端口号，以及源和目的IP地址的地理坐标（纬度和经度）。

特征	Mean	Std	25%	50%	75%
hour	11.41	7.19	6	11	18
day	10.18	0.60	10	10	11
源端口	48603.54	15328.04	38833.75	55630	58804
源纬度	30.33	18.47	22.26	37.75	37.75
源经度	19.53	86.91	-97.82	25.58	103.97
目标端口	2693.48	3933.13	135	445	5060
目标纬度	25.01	20.12	18.52	18.52	41.60
目标经度	1.26	73.32	-93.61	28.06	73.85



(a)



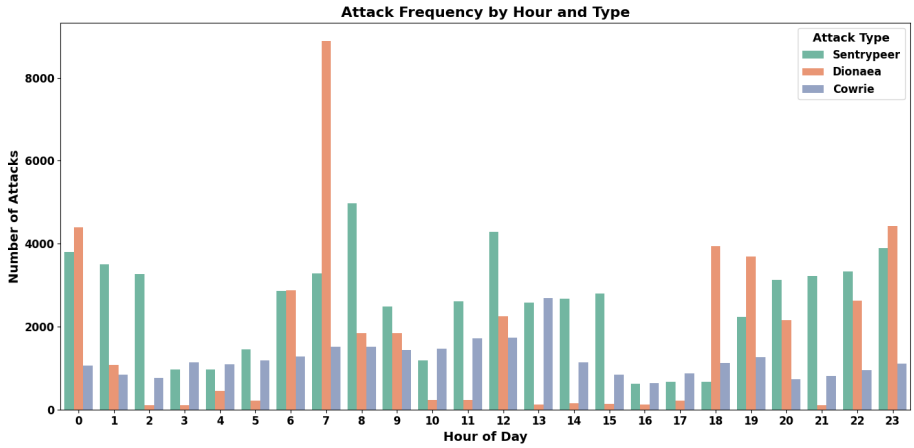
(b)

图2：网络攻击的时间模式。(a) 随时间变化的攻击类型演变，突出显示了 Diona 的峰值 ea 和 SentryPeer。(b) 攻击强度在各虚拟机上的分布，其中 vm-centralus 尤其成为攻击目标。

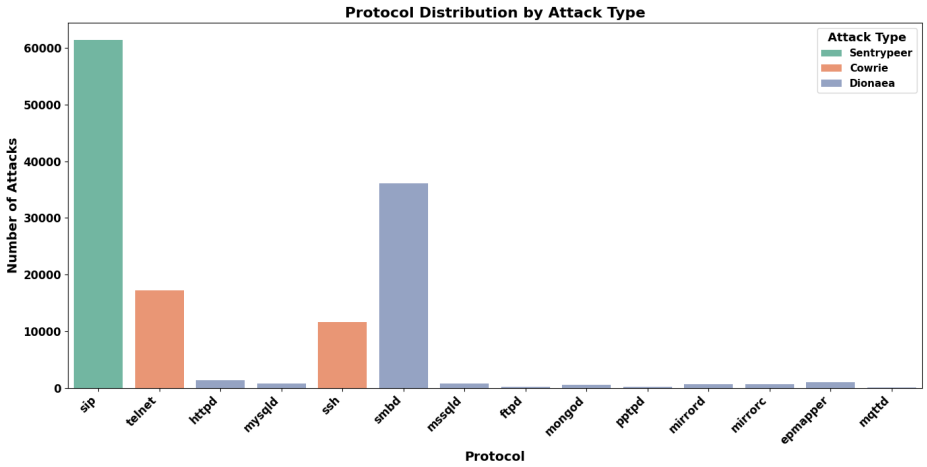
端点。这凸显了 Dionaea 专注于捕获针对 Windows 文件共享服务的恶意软件传播尝试。

HTTPD、MySQL、MSSQLD、MongoDB、PPTPD、Mirrord、MQTTP 和 EPMAP 等次要协议累计各自仅产生数千个事件，这表明攻击者在这些蜜罐上绝大多数都以三种主要攻击途径（SIP、Telnet 和 SMBD）为目标。不同蜜罐类型所呈现的协议偏好明显不同，这既反映了攻击者的行为，也凸显了各蜜罐服务仿真所带来的固有偏差。

图2(a)绘出了72小时窗口内按蜜罐类型细分的每小时攻击总量变化。出现了两种明显的锯齿状模式：首先，Dionaea（橙色）在6月9日协调世界时18:00左右出现尖峰，之后又在



(a)



(b)

图3：攻击模式的可视化。（a）每日趋势显示，攻击高峰出现在协调世界时 07:00 和 23:00。（b）各类蜜罐遭受滥用的协议，其中 SIP 尤其受到 SentryPeer 的攻击。

6月10日协调世界时23:00刚过再次出现；其次，SentryPeer（绿色）在6月10日和6月11日协调世界时约07:00均出现极高峰值。这些反复的升降并非仅由攻击者行为造成，而是直接对应于托管这些特定蜜罐的虚拟机按计划关闭和重启。每当虚拟机离线，其事件计数便会骤降至零；待其重新上线并恢复捕获探测后，计数又会再次上升。Cowrie（蓝色）也呈现出类似但不那么显著的波动模式，反映了其自身的维护周期。

图2(b) 展示了按虚拟机主机名汇总的攻击量，不区分蜜罐类型。这里的锯齿状波动更加明显：每台虚拟机——vm-centralus、vm-centralindia、vm-southafricanorth 和 vm-spaincentral——都出现间歇性低谷，攻击次数降至零，随后迅速恢复。值得注意的是，vm-centralus（橙色线）经历了最高峰值负载（单小时内超过 5,000 个事件），这凸显了它对攻击者的吸引力，或其更广泛的网络暴露程度。高频峰值与低谷交替出现，恰好对应虚拟机的运行计划，凸显了持续监控的重要性，以及维护时段可能造成监控盲区的风险。

图 4 展示了蜜罐遭遇活动的昼夜强度，其中每个单元格表示 Cowrie、Dionaea 和 SentryPeer 每小时记录的攻击次数。Dionaea 在协调世界时 07:00 出现明显且狭窄的峰值（8,885 个事件），成为所有平台中最繁忙的单个小时；这可能对应针对 Windows 文件共享服务（SMBD）的协同探测活动。相比之下，SentryPeer 在协调世界时 06:00 至 13:00 之间呈现持续的高活动平台期，在 07:00 达到峰值（3,121 个事件），并于 08:00 再次达到峰值（4,740 个事件），

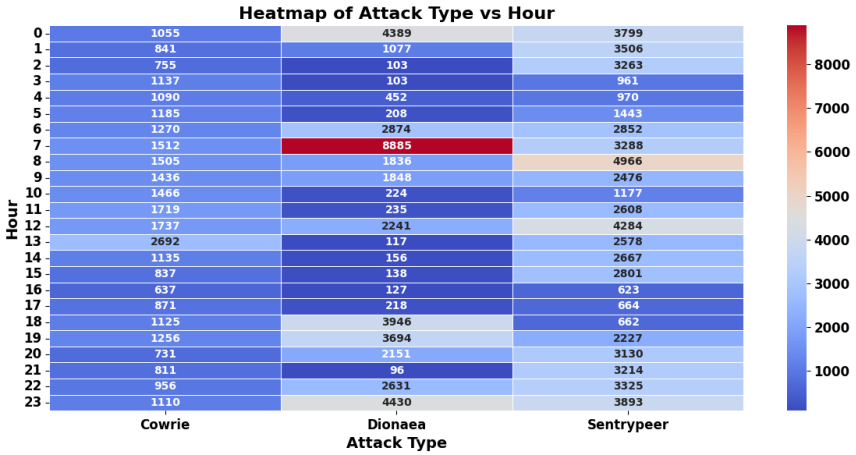


图 4：按小时和攻击类型显示攻击数量的热力图。Dionaea 的峰值出现在协调世界时 07:00。

反映了标准工作时间内持续针对 VoIP 的扫描活动。Cowrie 的 SSH/Telnet 流量相对均匀，每小时约有 700 至 1,700 次攻击，并在协调世界时 13:00 略有升高（2,692 个事件），表明远程访问服务全天候受到关注。三个蜜罐记录到的最低活动均出现在黎明前时段（协调世界时 02:00–04:00），这表明全球范围的自动化扫描有所减弱，或攻击者在这些时段的活动有所减少。这些各具特色的时间特征——Dionaea 短促而强烈的突发活动、SentryPeer 广泛的日间活动，以及 Cowrie 均匀的探测活动——凸显了时间感知防御措施的必要性，以便动态应对高峰时段激增和夜间低谷。

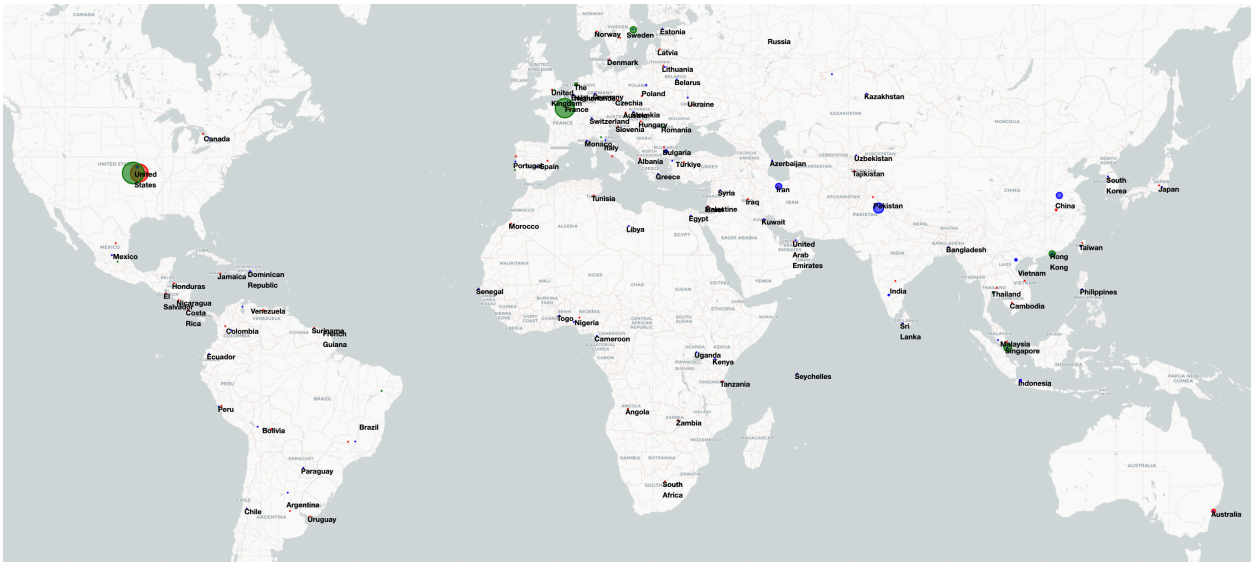


图 5：按蜜罐平台和来源国划分的网络攻击全球分布。圆圈大小与各国攻击总数成正比，颜色表示蜜罐类型：绿色（SentryPeer）、红色（Cowrie）和蓝色（Dionaea）。

图 5 展示了所有已记录入侵尝试的全球来源分布，并以圆圈直径表示各国的攻击量，以颜色区分涉及的蜜罐平台。绿色圆圈（SentryPeer）在美国及东南亚部分地区占主导，反映出大量面向SIP的探测。红色圆圈（Cowrie）在北美和西欧最为显著，表明 Telnet/SSH 扫描集中发生于这些地区。蓝色圆圈（Dionaea）聚集在欧洲各国，尤其是法国和英国，说明针对 SMB 的活动十分频繁。散布在南美洲、非洲和大洋洲的小型标记表明，全球扫描范围广泛，但强度较低

。这些空间分布模式表明，尽管攻击者遍布全球，但少数连接度较高的区域产生了大部分流量，这为防御性蜜罐部署的战略布局和扩展提供了依据。

记录与存储

该数据集被组织 为一个根目录和两个子文件夹：/RawData 和 /PreprocessedData。每个文件都 s 下文将对其进行说明，以便确保可复现性并便于导航。

根目录文件

根 目录包含用于说明数据集并定义预处理环境的文件：

- DataPreprocessingNotebook.ipynb：用于加载、清理和转换原始蜜网事件数据的 Jupyter 笔记本。其中包含用于解析 JSON 文件、处理缺失值、提取特征以及导出 CSV 文件的代码。
- README.md：纯文本文件，包含有关数据集及其结构的一般信息和使用说明。
- requirements.txt：列出了运行预处理笔记本所需的所有 Python 依赖项及其具体版本号，以确保结果可复现。

原始数据目录文件

The /RawData 文件夹包含未经处理的蜜网事件原始日志，采用 JSON 格式：

- RawHoneynetEventsBatch-0.json：第一批原始蜜罐遥测数据。
- RawHoneynetEventsBatch-1.json：第二批原始蜜罐遥测数据。
- RawHoneynetEventsBatch-2.json：第三批也是最后一批原始蜜罐遥测数据。

每个 JSON 文件都包含蜜罐交互的结构化记录，包括时间戳、源IP地址、目标端口和协议类型。

预处理数据目录中的文件

The /PreprocessedData 文件夹包含一个从原始 JSON 文件生成的清理后数据集 s:

- HoneyNetEvents_Clean.csv：所有原始批次汇总并预处理后的版本，格式为 CSV 文件。此数据集包含预处理笔记本中定义的提取特征。

预处理后的 CSV 汇总了全部三个原始 JSON 批次，并保留每个事件占一行的约定。所有特征均可根据原始文件和提供的笔记本重现。

见解与备注

研究人员应注意，该数据集存在若干局限与机遇。虚拟机重启导致的维护间隙会在时间序列分析中形成盲区，必须谨慎处理，可通过插值或明确建模为缺失数据来应对。每种蜜罐平台都存在固有偏差：SentryPeer 捕获的 SIP 洪泛攻击主要来自北美和东南亚，Cowrie 记录的 Telnet/SSH 扫描大多来自西欧和美国，而 Dionaea 记录的 SMB 漏洞利用则发生在欧洲节点周边。因此，在未考虑平台特定特征的情况下，不应将结果泛化。攻击活动高度集中，少数 IP地址产生了不成比例的大量流量，这凸显了采用适用于重尾分布的稳健统计方法的必要性。四区域部署也支持对攻击模式进行跨区域分析，但观察到的差异可能部分反映了当地互联网基础设施，而非纯粹的地理意图。除主要用于威胁分析外，该数据集还可用于异常检测算法基准测试、研究事件流中的时间突发性，以及探索真实环境中的协议滥用等次级应用，从而拓展其在原研究范围之外的价值。为支持可复现性，数据集保留了收集到的公网 IP 地址作为标识符；未包含任何其他个人数据。

源代码与脚本

完整的基础设施部署脚本、数据收集工具和分析代码均可在相关代码仓库中公开获取。该代码仓库提供了详细文档，介绍如何复现实验环境并扩展数据收集方法，其中包括：基础设施即代码模板、蜜罐配置文件、数据处理管道和分析脚本 [Feito-Casares 等人， 2025] [Feito Casares 等人， 2025]。所有代码均依据 MIT 许可证发布，以便学术和商业领域重复使用。

致谢与利益声明

本研究得到了 CyberFold 项目的资助。该项目由欧盟通过 NextGenerationEU 工具（复苏、转型与韧性计划）提供资金，并由西班牙国家网络安全研究所（INCIBE）管理，项目编号为 ETD202300129。本研究还得到了马德里自治区（ELLIS 马德里节点）以及西班牙科学与创新部通过项目 PID2022-140786NB-C32/AEI/10.13039/501100011033（LATENTIA）的资助。

E.F.C. 和 I.G.T. 整理并分析了数据，设计了实验基础设施，并撰写了稿件。J.L.R.A. 负责监督研究、提供方法指导并审阅稿件。所有作者均参与了最终稿件的构思与审阅。

作者声明不存在利益冲突。我们承认使用了人工智能助手（包括大语言模型）来辅助完成本研究。使用这些工具的任务包括校对、提升表达清晰度以及优化部分章节的结构。所有人工智能生成的建议均经作者审阅和批准，以确保学术诚信和准确性。

参考文献

Shishir Kumar Shandilya、Chirag Ganguli、Ivan Izonin 和 Atulya Kumar Nagar。用于深度数据包检测与分析的网络攻击评估数据集。*Data in Brief*，46:108771，2023年。Pedro Merino Laso、David Brosset 和 John Puentes。信息物理子系统中的异常与恶意行为数据集。*Datainbrief*，14:186–191，2017年。Daniel Tovarˇnák、Stanislav Špaˇcek 和 Jan Vykopal。网络防御演习期间采集的流量与日志数据。*Datain Brief*，31:105784，2020年。Ahmad T Al-Hammouri 和 Rasmus L Olsen。用于研究 Nagle 算法与延迟确认算法交互导致的信息物理系统通信缓冲延迟的数据集。*Data in brief*，39: 107530，2021年。Syaifuddin Saif、Widyawan Widyawan 和 Ridi Ferdiana。用于物联网中分布式拒绝服务攻击分类、识别与检测的 IoT-DH 数据集。*Data in Brief*，54:110496，2024年。Muhammad Faheem 和 Rizwan Aslam Butt。用于优化物联网赋能的工业 4.0 制造服务的大型光无线信息物理系统数据集。*Data in Brief*，42:108026，2022年。Verónica Valeros 和 Sebastián García。Hornet 40：地理分布式蜜罐网络数据集。*Data in Brief*，40:107795，2022年。Verónica Valeros 和 Sebastián García。CTU Hornet 65 Niner：地理分布式低交互蜜罐网络数据集。*Data in Brief*，58:111261，2025年。Markus Ring、Sarah Wunderlich、Deniz Scheuring、Dieter Landes 和 Andreas Hotho。基于网络的入侵检测数据集综述。*Computers & security*，86:147–167，2019年。Enrique Feito Casares、Ismael Gómez-Talal 和 José Luis Rojo-Álvarez。多区域蜜网数据集：跨 Azure 区域部署的分布式 T-Pot 蜜罐，用于网络威胁分析，2025年7月。Telekom Security。T-Pot 社区版。https://github.com/telekom-security/tpotce，2024年。访问日期：2025-06-22。Michel Oosterhof。Cowrie SSH/Telnet 蜜罐。https://github.com/cowrie/cowrie，2024年。访问日期：2025- 06-22。DinoTools。Dionaea：恶意软件陷阱。https://github.com/DinoTools/dionaea，2021年。访问日期：2025-06- 22。SentryPeer。SentryPeer SIP 蜜罐。https://github.com/SentryPeer/SentryPeer，2024年。访问日期：2025- 06-22。

Enrique Feito-Casares、Ismael Gómez-Talal 和 José Luis Rojo-Álvarez。云蜜网教程：用于在 Microsoft Azure 上收集威胁遥测数据的可扩展云原生 T-Pot 蜜罐基础设施，2025年7月。