

---

**《课题一：面向7种高隐蔽网络公害的网络侧分析模型构建  
及数据集调研报告》**

---

2025/06/30

远控  
木马

勒索  
病毒

漏洞  
利用

间谍  
软件

网络  
钓鱼

虚假  
流量

僵尸  
网络



序号	对象	类别	名称
1	恶意软件	源码	VX underground Malware Source Code
2	恶意工具代码	源码	Qakbot
3	恶意网络流量	数据集	Stratosphere IPS Datasets
4	真实大型网络流量	数据集	MAWI-2020 Datasets
5	DDoS流量	数据集	CIC-DDoS-2019 Datasets
6	入侵检测流量	数据集	CIDDS (Coburg Intrusion Detection Data Sets)
7	恶意软件流量	数据集	MTA-KDD19
8	DoH流量/恶意DoH流量	数据集	CIRA-CIC-DoHBrw-2020 Datasets
9	QUIC协议流量	数据集	CESNET-QUIC22
10	车联网入侵检测流量	数据集	CIC-IoV-2024
11	恶意软件侧信道数据	数据集	side-channel-data-malware-intrusion-detection
12	恶意流量/恶意流量攻击指令	数据集	DataCon Datasets
13	恶意软件流量	论文 (网站停运)	Using side channel TCP features for real-time detection of malware connections
14	恶意软件样本	IOC情报	MalwareBazaar
15	恶意软件样本	IOC情报	Malware-IOCs
16	恶意软件样本	IOC情报	VirusShare
17	恶意SSL证书/恶意软件指纹	IOC情报	SSL Blacklist (SSL BL)

序号	对象	类别	名称
1	恶意软件	源码	VX underground Malware Source Code
2	恶意工具代码	源码	Qakbot

远控木马 ✕

勒索病毒 ✕

漏洞利用 ✕

间谍软件 ✕

网络钓鱼 ✕

虚假流量 ✕

僵尸网络 ✕

## VX ug MSC / Qakbot

## 概述

- VX Underground是一个集合了各种恶意软件源代码的仓库
- Qakbot是一个包含Qakbot木马核心代码的仓库，包括其网络通信模块的实现。

## 适用性分析

- 两者都只提供源代码本身，并不直接涉及网络流量或协议数据，因此不存在“是否包含加密流量协议”的说法。
- VX ug理论上涉及所有七类高公害的源码，尤其含有 **木马 / 勒索 / 间谍** 等完整通信实现。
- Qakbot只包含 **木马 / 勒索** 这个类别。

## 总结

- 需 **编译+沙箱** 才能抓取流量，**难度较高、风险较高**，且**捕获的加密恶意流量质量可能难以保证**。
- 只适合 **协议指纹 & C2 行为 & 证书指纹** 逆向分析

序号	对象	类别	名称
3	恶意网络流量	数据集	Stratosphere IPS Datasets
4	真实大型网络流量	数据集	MAWI-2020 Datasets

- 远控木马 ✓
- 勒索病毒 ☹️
- 漏洞利用 ☹️
- 间谍软件 ☹️
- 网络钓鱼 ✖️
- 虚假流量 ✖️
- 僵尸网络 ✓

Stratosphere IPS Datasets

概述

- 2011-2020年，新老均有（新的如IoT-23）；
- 真实的**恶意** + **良性PCAP**，但混合着包含加密和未加密的恶意流量，用前需**筛选出加密流量**；
- 有**完整详细的标签**：如僵尸网络、RAT、恶意扫描

适用性分析

- IoT-23/CTU-13中的是**大量僵尸网络流量**，一部分**远控木马/RAT流量**，以及**少量扫描和漏洞利用行为**（如CTU-13中部分场景涉及端口扫描等）。
- 僵尸网络木马具备信息窃取功能（可视作**间谍行为**）或充当勒索病毒载体。

总结

- **优点** 真实且带标签，方便网络侧训练检测模型；
- **缺点** 混合着未加密流量，训练前需筛选
- 主要适用于**僵尸网络**和**远控木马**检测。
- 可能适用于**勒索 / 间谍 / 漏洞**

MAWI-2020 Datasets

概述

- 日本骨干网15 min/天抓包，HTTPS/TLS加密流量占比高
- **无标签** ⇒ **高真实性背景**，适合用来压测误报
- **数据规模极大**，需采样处理

适用性分析

- 对于本课题，MAWI数据集本身**不直接提供恶意流量样本**，一般当作**良性背景流量**与已知**恶意流量**混合使用。
- 可能混杂少量恶意流量（如扫描、攻击等背景出现，但未标注），根据具体情况决定是否忽略。

总结

- MAWI本身不针对特定威胁类别，主要扮演**良性流量背景角色**。
- **七种高隐蔽网络公害类别均可使用**，但只能**辅助使用**（数据增强，提高鲁棒性）。

序号	对象	类别	名称
5	DDoS流量	数据集	CIC-DDoS-2019 Datasets
6	入侵检测流量	数据集	CIDDS (Coburg Intrusion Detection Data Sets)

- 远控木马 ✓
- 勒索病毒 ✗
- 漏洞利用 ✓
- 间谍软件 ✗
- 网络钓鱼 ✗
- 虚假流量 ✓
- 僵尸网络 ✓

CIC-DDoS-2019

概述

- 多种DDoS流量 + 正常流量；
- 包含加密流量；
- PCAP & CSV，pcap源文件方便我们自行处理

适用性分析

- 僵尸网络经常用于发动DDoS攻击，可以算作僵尸网络的活动危害。
- 虚假流量类别的威胁——攻击者产生海量恶意流量扰乱服务。
- 其余5种类别关联不大

总结

- 主要对应僵尸网络/虚假流量类别。
- 对课题可用。

CIDDS-001 (NetFlow)

概述

- 仿真企业网，提供 NetFlow 形式的数据集（无载荷，每条记录只包含通信两端IP/端口、协议、字节数、数据包数、持续时间等摘要信息）。
- 包括扫描、DDoS、内部渗透等多种攻击行为流量

适用性分析

- 数据以NetFlow形式CSV给出，没有原始PCAP。
- NetFlow形式 ⇒ 没有应用层载荷，只有元数据特征（通信两端IP/端口、协议、字节数、数据包数、持续时间等摘要信息）
- 因此无论流量是否加密都不影响，适合加密恶意流量分析中“加密”的特性。

总结

- 涵盖远控木马（内部恶意主机外联）、虚假流量/DDoS（有模拟攻击流量）、漏洞利用（扫描、内部渗透等攻击行为）。
- 对课题可用。

序号	对象	类别	名称
7	恶意软件流量	数据集	MTA-KDD19
8	DoH流量/恶意DoH流量	数据集	CIRA-CIC-DoHBrw-2020 Datasets

- 远控木马 ✓
- 勒索病毒 ✗
- 漏洞利用 ✗
- 间谍软件 ✓
- 网络钓鱼 ✗
- 虚假流量 ✗
- 僵尸网络 ✓

MTA-KDD'19

概述

- 约7.15万条流的特征记录，其中55.3%恶意、44.7%良性；
- 每条记录包含33维特征，涵盖了各种统计指标（如包间隔、包大小分布、TCP标志计数、DNS请求数等）

适用性分析

- 与CIDDS-001（NetFlow）
- 数据以NetFlow形式CSV给出，没有原始PCAP。
- NetFlow形式 ⇒ 没有应用层载荷，只有元数据特征（通信两端IP/端口、协议、字节数、数据包数、持续时间等摘要信息）
- 因此无论流量是否加密都不影响，适合加密恶意流量分析中“加密”的特性。

总结

- 主要对应的类别是：远控木马 / 僵尸网络
- 是加密的流量的数据集
- 对课题可用

CIRA-CIC-DoHBrw-2020

概述

- 100% HTTPS(DoH)，恶意请求占 ≈ 90 %
- 协议上全部是加密的HTTPS流量，符合“加密”要求
- 约37万条流记录每条流记录提取了相应的特征，类似CIC其他数据集风格，有CSV格式。

适用性分析

- 直接关联间谍软件和远控木马类别（因为使用DoH隧道的多是渗透者用于隐藏其控制通信或数据窃取）。
- 与虚假流量也有一点相关——恶意流量伪装成正常的Web加密流量正是一种虚假/伪装行为，但联系没那么强

总结

- 主要对应的类别是：远控木马 / 间谍软件
- 是加密的流量的数据集
- 对课题可用

序号	对象	类别	名称
9	QUIC协议流量	数据集	CESNET-QUIC22
12	恶意流量/恶意流量攻击指令	数据集	DataCon Datasets

- 远控木马 ✓
- 勒索病毒 ✗
- 漏洞利用 ✗
- 间谍软件 ✓
- 网络钓鱼 ✗
- 虚假流量 ✗
- 僵尸网络 ✗

CESNET-QUIC22

概述

- ISP 骨干网 1.5 亿条 QUIC 加密流 (2022) ;
- 1.53亿条QUIC流 (约89GB) , 提供元数据特征
- 每条流的元数据特征包括: 字节数、包数、持续时间以及QUIC特有的信息 (如TLS握手中的SNI域名、User-Agent字符串、QUIC版本) , 还记录了每条流前30个数据包的大小和时间间隔序列
- 没有直接的“攻击”或“恶意”标注。

适用性分析

- 没有直接标签说明哪些流是恶意, 但可以利用已知IOC去匹配: 例如根据SSL证书指纹或SNI域名, 将可能属于已知恶意域名的QUIC流标记出来进行分析 (如结合SSL BlackList这个IOC情报库)
- 或者只是单纯当背景流量。
- CESNET-QUIC22本身不区分威胁类型, 但未来各种威胁都可能迁移到QUIC之上 (前瞻)

总结

- 如果利用IOC标记, 可能捕捉到僵尸网络C2或远控木马的QUIC通信

DataCon 2020 & 2021

概述

- Cobalt Strike或类似远控工具的恶意流量, 混合着正常流量
- 提供pcap, 均是加密https流量

适用性分析

- DataCon的数据还涉及识别具体“攻击指令”, 标注了更细的类别 (例如C2心跳包/C2命令包) , 有待继续调研。
- 理论上也可以只提取常规的元特征进行检测, 如TLS证书特征、流量包长统计、流量时间序列特征等。

总结

- 非常对应远控木马 / 间谍软件。
- 包含真实加密恶意通信的流量数据, 对课题有用。



序号	对象	类别	名称
10	车联网入侵检测流量	数据集	CIC-IoV-2024
11	恶意软件侧信道数据	数据集	side-channel-data-malware-intrusion-detection
13	恶意软件流量	论文（网站停运）	Using side channel TCP features for real-time detection of malware connections

- 远控木马
- 勒索病毒
- 漏洞利用
- 间谍软件
- 网络钓鱼
- 虚假流量
- 僵尸网络

CIC IoV-2024

概述

- 2024年
- 完全是汽车内部的CAN总线协议数据，而非典型的IP网络流量，也非加密；

适用性分析

- 与课题聚焦的 网络侧 加密 检测方向不匹配

总结

- 与课题聚焦的 网络侧 加密 检测方向不匹配

硬件侧信道恶意软件检测数据集

概述

- 2020年
- 包含在多台计算机上运行恶意软件时采集的各种硬件性能指标时间序列。
- CPU各核心温度、CPU利用率、内存占用率、磁盘I/O、风扇转速等

适用性分析

- 不涉及任何网络流量或协议，属于端侧
- 与课题聚焦的 网络侧 加密 检测方向不匹配

总结

- 价值：启发侧信道概念；实际网侧模型不可用

TCP侧信道特征论文

概述

- 2019年
- 中科院四区 JCR三区
- 提供原始数据集或代码的网站停运关闭，无法获取

适用性分析

- 不解析加密流量内容，仅通过侧信道特征包括诸如TCP报文的大小、时间间隔、序列模式等
- 特征设计的思路可参考借鉴

总结

- 原始数据集或代码不可用，需要自行找数据集并复现



序号	对象	类别	名称
14	恶意软件样本	IOC情报	MalwareBazaar
15	恶意软件样本	IOC情报	Malware-IOCs
16	恶意软件样本	IOC情报	VirusShare
17	恶意SSL证书/恶意软件指纹	IOC情报	SSL Blacklist (SSL BL)

远控木马 ☹️

勒索病毒 ☹️

漏洞利用 ❌

间谍软件 ☹️

网络钓鱼 ❌

虚假流量 ❌

僵尸网络 ☹️

### Malware-IOCs GitHub

- 每日更新，域名 / IP / 哈希
- 覆盖 RAT / 间谍 / 钓鱼 / 僵尸 等
- 可用于 自动标注流量 & 规则拦截
- 局限：仅识别已知威胁，需与模型互补

### MalwareBazaar

- 样本下载
- 能够快速获取最新的 勒索病毒 / 间谍软件的样本，以及C2列表

### VirusShare

- 与MalwareBazaar类似，提供样本下载服务

### SSL Blacklist (SSLBL)

- 提供网侧威胁的 恶意证书指纹 & JA3/JA3S（持续更新黑名单列表）
- 可作为模型的“初步筛选阶段”用途，开销小。（辅助作用）
- 覆盖：木马 / 间谍 / 勒索 / 僵尸 等

序号	对象	类别	名称	远控 木马	勒索 病毒	漏洞 利用	间谍 软件	网络 钓鱼	虚假 流量	僵尸 网络	*备注
1	恶意软件	源码	VX underground Malware								
2	恶意工具代码	源码	pr0xylife/Qakbot								
3	恶意网络流量	数据集	Stratosphere IPS Datasets	✓						✓	
4	真实大型网络流量	数据集	MAWI-2020 Datasets	—	—	—	—	—	—	—	*背景良性流量
5	DDoS流量	数据集	CIC-DDoS-2019 Datasets						✓	✓	
6	入侵检测流量	数据集	CIDDS-001	✓		✓			✓		
7	恶意软件流量	数据集	MTA-KDD19	✓						✓	
8	DoH流量/恶意DoH流量	数据集	CIRA-CIC-DoHBrw-2020 Datasets	✓			✓				
9	QUIC协议流量	数据集	CESNET-QUIC22	✓						✓	*QUIC下的通信
12	恶意流量/恶意流量攻击指令	数据集	DataCon Datasets	✓			✓				
10	车联网入侵检测流量	数据集	CIC-IoV-2024								
11	恶意软件侧信道数据	数据集	Side-channel-data-malware-intrusion-detection								
13	恶意软件流量	论文（网站停运）	Using side channel TCP features ...								
14	恶意软件样本	IOC情报	MalwareBazaar	—	—	—	—	—		—	*需与模型互补，仅辅助
15	恶意软件样本	IOC情报	Malware-IOCs								
16	恶意软件样本	IOC情报	VirusShare								
17	恶意SSL证书/恶意软件指纹	IOC情报	SSL Blacklist (SSL BL)	—	—		—			—	*用作预筛选，仅辅助