软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.007022]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于突变流量的在野黑产应用采集方法 *

陈沛[1], 洪赓[1], 邬梦莹[1], 陈晋松[1], 段海新[2,3], 杨珉[1]

[1](复旦大学 计算机科学技术学院, 上海 201203)

[2](清华大学 网络科学与网络空间研究院, 北京 100084)

[3](中关村实验室, 北京 100081)

通讯作者: 杨珉, E-mail: m_yang@fudan.edu.cn

摘 要: 随着经济社会的快速发展, 互联网黑色产业(也称互联网地下产业, 以下简称网络黑产)对人民群众的生产生活带来的影响也在快速扩大. 近年来, 移动互联网的兴起使以诈骗、博彩和色情为主的网络黑产移动应用(APP)变得更加猖獗, 亟待采取有效措施进行管控. 目前研究人员针对黑产应用的研究较少, 其原因是由于执法部门持续对传统黑产应用分发渠道的打击, 已有的通过基于搜索引擎和应用商店的采集方法的效果不佳, 缺乏大规模具有代表性的在野黑产应用数据集已经成为开展深入研究的一大掣肘. 为此, 本文尝试解决在野黑产应用大规模采集的难题, 为后续深入全面分析黑产应用及其生态提供数据支撑. 本文提出了一种基于突变流量分析的黑产应用批量捕获方法, 以黑产应用分发的关键途径为抓手, 利用其具有的突变和伴随流量特点, 批量快速发现正处于传播阶段的新兴在野黑产应用, 为后续实时分析和追踪提供数据基础. 在测试中, 本方法成功获取了 3,439 条应用下载链接和 3,303 个不同的应用. 捕获的移动应用中, 不但有 91.61%的样本被标记为恶意软件, 更有 98.14%的样本为首次采集发现的零天应用. 上述结果证明了本文提出的方法在黑产应用采集方面的有效性.

关键词: 互联网地下产业;网络黑产;移动应用;流量分析

中图法分类号: TP311

中文引用格式: 陈沛,洪赓,邬梦莹,陈晋松,段海新,杨珉.一种基于突变流量的在野黑产应用采集方法.软件学报. http://www.jos.org.cn/1000-9825/7022.htm

英文引用格式: Chen P, Hong G, Wu MY, Chen JS, Duan HX, Yang M. An underground industry application collection method based on flow analysis. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7022.htm

An underground industry application collection method based on flow analysis

CHEN Pei[1], HONG Geng[1], WU Meng-Ying[1], CHEN Jin-Song[1], DUAN Hai-Xin[2,3], YANG Min[1]

[1](School of Computer Science, Fudan University, Shanghai 201203, China)

[2](Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China)

[3](Zhongguancun Lab, Beijing 100081, China)

Abstract: In recent years, with the rise of the mobile Internet, underground mobile applications primarily involved in scams, gambling, and pornography have become more rampant, requiring effective control measures. Currently, there is a lack of research on underground

* 基金项目: 国家自然科学基金(62302101)

陈沛和洪赓为共同第一作者, 作者顺序依据姓氏首字母排序.

收稿时间: 2023-09-11; 修改时间: 2023-10-30; 采用时间: 2023-12-15; jos 在线出版时间: 2024-01-05

applications by researchers. Due to the continuous crackdown by law enforcement agencies on traditional distribution channels for these applications, the existing collection methods based on search engines and app stores have proven to be ineffective. The lack of large-scale and representative datasets of real-world underground applications has become a major constraint for in-depth research. Therefore, our paper aims to address the challenge of collection of large-scale real-world underground applications, providing data support for a comprehensive in-depth analysis of these applications and their ecosystem. We propose a method to capture underground applications based on traffic analysis. By focusing on the key distribution channels of underground applications and leveraging their characteristics of mutation and accompanying traffic, we can discover in-the-wild underground applications in the propagation stage. In the test, this method successfully obtained 3,439 application download links and 3,303 distinct applications. Among the apps, 91.61% of the samples were labeled as malware by antivirus engine, while 98.14% of the samples were zero-days. The results demonstrate the effectiveness of our proposed method in the collection of underground applications.

Key words: underground ecosystem; underground app market; mobile apps; traffic analysis

随着经济社会快速发展, 互联网黑色产业(以下简称网络黑产)的犯罪手段和范围也迅速发展. 近年来, 随着移动互联网的兴起, 以诈骗、博彩、色情为代表的网络黑产移动应用(APP)日渐猖獗, 严重影响人民群众生产生活. 360 数字安全集团发布的《2022 年度反诈报告》显示, 近一年新增了 2400 万恶意应用, 日均新增 6.5 万, 其中主要以投资理财、色情和赌博三类为主[1]. 随着人民生产生活对于移动应用日渐依赖, 传统意义的犯罪形式进一步向移动端迁移. 公安部数据表明, 截止 2021 年, 通过诈骗 APP 实施的电信网络诈骗发案已经成为电信网络诈骗案的主要形式, 占比超过 60%[2]. 因此, 亟需采取措施对黑产应用加以管控.

当前, 已有对于网络黑产攻击方法研究主要关注于黑产网站检测方面, 网站数据多来自于搜索引擎和公开数据集. Yang 等人[3]通过爬取合作搜索引擎获得了其索引的大量网站候选域名, 使用自然语言处理技术过滤识别大量非法博彩网站, 并对博彩网站的推广策略、第三方支付渠道、网络存储服务等进行了研究. 此外, 部分研究者使用内容审核领域中的公开图片或视频数据集, 结合文本及图像特征采用机器学习方法对敏感内容进行检测[4-8]. 除了被动检测外, 网络空间资产测绘引擎 Shodan[9]、Zoomeye[10]等通过扫描互联网 IPv4 地址并匹配黑产指纹特征, 主动探测黑产网站.

已有研究中, 对于黑产应用的研究还较为匮乏, 其首要掣肘于缺乏鲜活的在野黑产应用数据集. 当前, 大多数研究的黑产应用样本需要研究者通过搜索引擎收集或需要研究者与有关部门建立联系. Gao 等人[11]通过爬取搜索引擎结果中的博彩网站内嵌入的博彩应用, 系统地分析了博彩应用的生态和非法博彩应用程序之间的联系. 这种基于在搜索引擎利用关键词搜索并过滤的收集方式依赖于搜索关键词, 一方面局限于对于某些已知特定领域, 且由于搜索引擎存在安全检测机制, 大多数如博彩网站在内的违法网站并不会被搜索引擎所收录, 因此这种应用收集方法效率较低. Hong 等人[12]基于匿名机构提供的 1,487 个赌博应用程序开展了博彩诈骗的实证研究, 此种研究方法收集应用程序的丰富度和效率较低, 在样本采集的扩展性上较为受限.

本文希望解决黑产应用的大规模采集问题, 为后续开展黑产应用深入研究奠定良好基础. 但在野黑产应用的批量采集并非易事, 当前黑产应用分发传播往往呈现出高对抗性, 这为大规模采集工作带来了诸多挑战: 首先, 黑产应用为了躲避监管多选择隐匿传播, 因此不会在常规应用商店中上架推广. 其次, 黑产团伙在自身网站中推广少量黑产应用, 但黑产网站数量级远远超过其推广的黑产应用, 通过黑产网站采集方案的效率低下.

针对现有黑产应用收集难题, 本文提出了一种基于突变流量分析的黑产应用发现方法. 该方法不依靠单个黑产网站或人工收集, 而是基于黑产应用分发中的关键途径对黑产应用进行批量捕获. 该方法基于以下观察:

（1） 随着正规应用商店(如华为应用市场、360 应用商店)中批量取缔下架黑产应用, 当前, 在黑产应用往往通过一类特殊的网站, 分发黑产应用, 我们称作黑产门户. 黑产门户网站往往提供

了大量黑产应用的下载链接, 部分黑产门户同时为多个黑产团伙应用提供下载服务. 同时, 黑产门户网站还会定期更换应用下载链接, 从而保证下载应用的有效性. 因此, 黑产门户可以作为下载新鲜、正处于传播过程中的黑产应用的渠道.

（2） 黑产门户网站除了分发黑产应用, 通常还添加多种黑产网站广告链接、为各类黑产网站引流来获取利润. 首先, 由于黑产门户网站本身具有较大流量, 被引流网站在引流前后的访问量量级差距明显, 存在访问量突变. 其次, 用户在连续访问黑产门户和被引流黑产网站时, 其访问序列表现出时序临近关系. 因此, 域名流量突变特征可以作为批量发现黑产门户网站的重要思路.

本文提出的方法利用黑产门户网站作为黑产应用采集来源, 可有效地解决传统应用收集方式存在的问题. 首先, 基于被动 DNS 数据库计算获取发生访问量突变的域名, 它们是潜在的被引流对象. 然后, 通过伴随算法反查访问者在临近时间内的访问序列, 得到的网站可能就是为突变域名引流的门户网站, 最后, 本文构建了基于多种特征的二分类模型, 通过模型判别得到黑产门户, 通过这些黑产门户, 便可以下载到应用. 基于传播渠道的方案存在以下优点: 采集数量大、效率高, 从一个黑产门户中往往可以采集到大量应用; 数据来源鲜活, 有助于找到正处于分发阶段的在野黑产应用, 更具分析与追踪价值.

我们使用 2023 年 5 月的被动 DNS 数据对该方法的发现能力进行了评估. 经测试, 累计获取 3,439 条应用下载链接和 3,303 个不同的移动应用. 为了评估这些应用的类型, 我们使用 VirusTotal[13]对下载到的应用进行检测, 发现其中 91.61%的应用被至少一个引擎检出为恶意软件, 98.14%的应用为零天应用, 证明了本文方法可有效发现黑产应用.

本文第 1 节介绍本文使用到的相关技术的基础知识和概念, 包括被动 DNS 技术和黑产门户的概念. 第 2 节介绍本文构建的基于突变流量分析的黑产应用发现方法. 第 3 节通过真实数据测试实验验证所提算法的有效性, 并对本方法进行评估. 第 4 节介绍网络黑色产业链的相关研究现状. 最后第 5 节总结全文.

1 背景知识

本文所提方法主要基于被动 DNS 和黑产门户概念, 本节就相关概念和基本知识予以介绍.

1.1 被动DNS原理

被动 DNS(Passive DNS)是一种用于收集和记录域名系统(DNS)活动的技术和方法[14-15]. 与主动 DNS 通过用户向 DNS 服务器查询域名解析信息不同, 被动 DNS 将全球域名系统中可用的 DNS 数据信息重建到中央数据库中, 以便研究人员对 DNS 解析进行反向检索和查询.

被动 DNS 的工作原理如下:

(1) 数据收集: 被动 DNS 服务器监听网络中的 DNS 查询, 并返回相应的响应数据包.

(2) 数据记录: 收集到的 DNS 查询和响应数据包会被记录下来, 并保存为结构化的数据. 通常, 被动 DNS 服务器将这些数据存储在数据库中, 以便后续的查询和分析.

(3) 数据分析: 通过对被动 DNS 数据进行分析, 可以获得有关域名的详细信息, 例如域名的解析历史、IP 地址变化、域名关联关系等. 此类数据往往用于网络安全监测、威胁情报分析、恶意域名检测等方面.

在本研究中, 被动 DNS 数据项可以被形式化地概括为三元组

$$< IP, domain, timestamp > \tag{1}$$

该三元组的含义为, 某一个客户端 IP 在某一时刻(timestamp)发出了一个 DNS 查询请求, 该请求内容为查询 domain 的实际地址. 当三元组数量足够大时, 通过聚合操作可以侧面表示更多用户信息. 例如, 对于 IP 相同的不同三元组, 此序列反映了该 IP 用户(或用户组)在不同时刻请求解析的域名及请求时刻, 这些时序行为数据可被用于分析用户行为; 对于 domain 相同的不同三元组, 此序列反映了该域名在被不同用户(或用户组)访问的时刻与次数, 此行为数据可被用于域名访问分析, 寻找潜在的恶意访问流量等.

值得注意的是, 出于对用户隐私保护的考量, 本文中所涉及的客户端 IP 均已匿名化处理, 不含任何个人识别信息(PII, Personally Identifiable Information).

1.2 网络黑产应用产业链

网络黑产是指利用互联网技术实施网络违法行为, 以及为这些行为提供工具、资源、平台等准备和非法获利变现的渠道与环节[16]. 黑产相关违法行为对个人和企业都会造成严重的影响. 意大利信息安全协会在最近一份报告中称, 仅2021年, 黑客攻击和各种网络犯罪就给全球经济造成了超过6万亿美元的损失, 预计到2025年, 此类犯罪每年将使全球经济损失约10.5万亿美元[17].

随着监管及执法机构对黑产网站的快速封禁打击, 网络黑产逐渐将犯罪活动转移到移动端平台上. 这些网络犯罪团伙通过批量制作、分发黑产应用, 将受害者引导至由黑产团伙所控制的应用环境内, 欺骗、诱导受害者完成大量交易, 最终转化为犯罪分子的金钱收益.

网络黑产应用产业链主要有以下参与者:

(1) 应用制作者: 制作黑产应用的开发者、后台控制团伙. 由于黑产应用快速上线的特性, 他们可能使用在线应用生成器(Online Application Generator, OAG) 或者私有的应用生成工具批量低成本制作黑产应用[12]. 在一些情况下, 应用制作者还可能会操控应用后台, 修改数据, 诱导受害者上当.

(2) 黑产门户: 批量分发、传播黑产应用的平台. 一方面, 黑产门户是用户直接接触黑产应用的窗口, 用户可以通过黑产门户浏览、选择和下载的黑产应用. 另一方面, 黑产门户是黑产应用的流量入口, 黑产应用及其制作者通过购买广告的方式上架黑产门户, 增加其影响力进而提高收入.

(3) 支付渠道: 为黑产应用中的交易提供支付鲁棒的收款服务. 在目前的监管政策下, 包括银行、第三方支付渠道(如支付宝、微信支付等)在内的各大支付渠道都禁止黑产应用使用相关服务. 因此出现了专门为黑产团伙提供的高度稳定和匿名的支付渠道及服务, 实现非法交易变现.

(4) 用户: 黑产应用的使用者、受害者. 用户通过黑产门户下载黑产应用, 并在使用应用的过程中受到犯罪团伙诱导投入个人资金, 或被黑产应用窃取敏感隐私, 导致移动设备无法正常使用等. 同时, 其金钱投入为整个黑色产业链提供源源不断的外部资金输入.

以上四类参与者在网络黑产应用产业链上的关系如图 1.

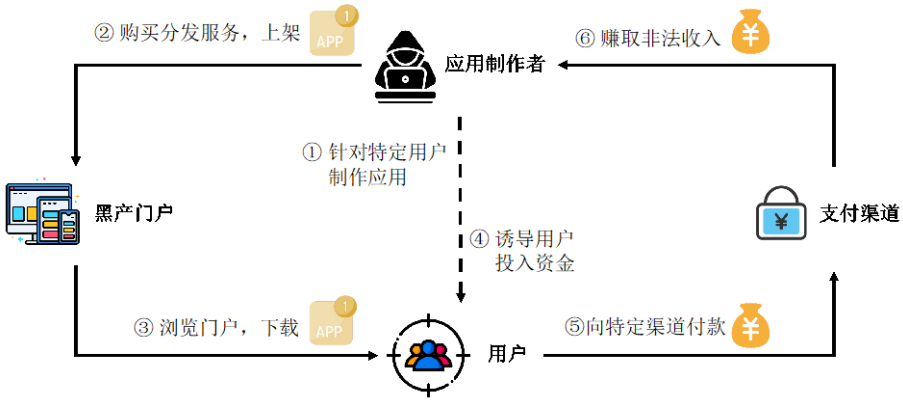


图 1 黑产应用产业链示意图(彩印)

2 基于突变流量分析的黑产应用发现方案设计

本节首先介绍我们提出的基于流量分析的黑产应用发现方法的核心思路和设计上的考虑, 其次介绍方案的总体框架流程, 以及发现流程中的 3 个主要步骤.

2.1 核心观察

发现并捕获在野黑产应用的关键在于掌握黑产应用的分发渠道. 对于该问题, 本文有以下两点核心观察:

- 观察一: 黑产门户是黑产应用的重要分发渠道

在黑产应用的分发过程中, 存在一类特殊的分发网站, 其往往提供了大量黑产应用的下载链接, 并为多个黑产应用提供分发服务, 我们将其称作黑产门户网站, 简称黑产门户. 同时, 黑产门户为保证下载应用的有效性, 其运营者会定期更新维护黑产应用下载链接. 利用该特点, 我们可以将黑产门户作为捕获在野应用的关键渠道.

图 2 展示了黑产门户示例. 如图, 这些网站界面与常规应用商店相似, 其中包含醒目的应用图标和诱导用户下载的对应按键. 此外, 部分黑产门户还可能带有页面广告, 通过给相关地下产业导流获利.

图 2 黑产门户示例(彩印)

值得注意的是, 黑产门户和传统的黑产网站存在较大的区别: 黑产网站包括网络黑产团伙所部署的赌博、色情、诈骗网站, 尽管部分黑产网站会通过直接在网站上推荐访问者下载来自同一开发集团的应用, 已达到部署分发应用的目的. 但此类网站通常只会推广单个应用, 因此无法作为稳定、批量的应用收集渠道. 黑产门户网站往往集成了包含多种类型、多个利益集团的黑产应用下载链接, 并以醒目的色彩诱导用户下载应用的网站, 是当前受害者下载黑产应用的最主要渠道之一. 同时, 黑产门户通常流量规模较大, 为通过分发应用获利, 其内嵌的应用下载渠道普遍具有较高的及时性和多样性, 因此若能对通过黑产门户传播的应用进行有效侦测, 则能有效捕获较多的在野黑产应用.

- 观察二: 黑产门户与其下游的网站的访问量存在显著关联特征

网站间的引流关系指的是不同网站之间通过直接嵌入广告跳转链接或间接文字提示建立起的流量引导关系. 它的主要表现为, 引流网站通常访问量较大、流量稳定, 而且被引流网站在被引流前通常访问量较低, 被引流后发生访问量突增. 在本文所研究的问题中, 黑产门户网站通常具有稳定的流量, 它们通常还添加多种黑产网站广告链接、为各类黑产网站引流来获取利润. 因此, 这种引流行为使得双方的流量关系可被建模.

本文进一步形式化地定义一种网站间的引流关系, 该关系建模目的为帮助定位黑产门户.

(1) 突变域名: 定义发生了访问量突变的被引流网站为突变域名. 黑产门户网站作为高流量平台, 向下游网站导流行为会为被引流网站带来大量用户访问, 从而对被引流网站的访问量产生显著

的影响. 这将导致被引流网站在此前后的访问量量级差距明显, 导致被引流网站出现访问量突然变化. 利用这一现象, 使得我们从突变现象出发寻找背后的黑产门户成为可能, 基于此, 可以进一步监测黑产活动, 从而采取相应的防范措施.

(2) 伴随域名: 定义与突变域名存在临近访问时序关系的域名集合为伴随域名. 引流关系的实质是, 用户在访问黑产门户后, 在临近时间内由黑产门户中的引流链接跳转到了被引流网站上. 这一操作使得用户的访问序列表现出时序临近关系, 在这个临近关系宏观上表现为被动 DNS 数据中的时间戳差值在一定阈值内. 基于这种现象, 可以设计伴随算法筛选满足访问时序临近的域名, 从而能进一步得到黑产门户候选集合, 为批量发现黑产门户网站、黑产应用提供数据来源. 根据以上定义, 可从与突变域名具有伴随关系的域名中筛选出黑产门户.

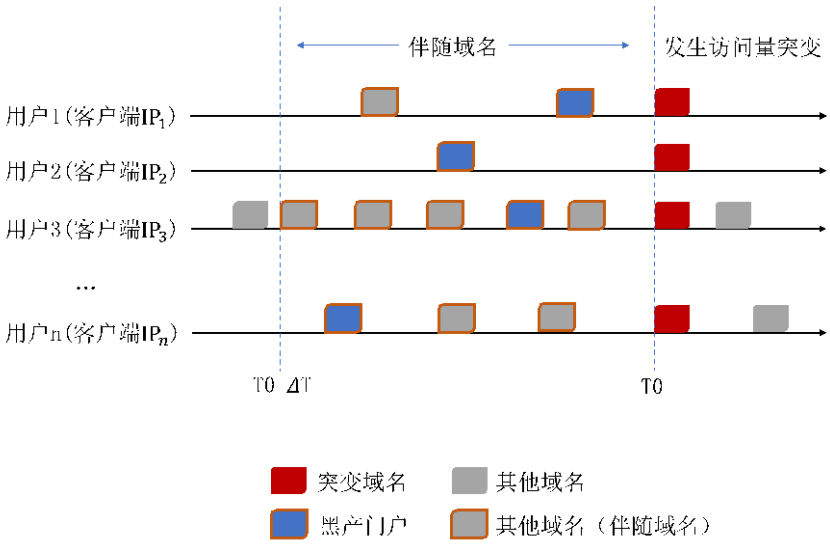


图 3 突变伴随关系示意图(彩印)

如图 3 所示, $IP_1, IP_2, ..., IP_n$ 为已知突变域名(图中红色网站), 在发生突变当天, 通过被动 DNS 数据反查到的所有访问 IP. 假设 IP_i 访问该突变域名的时间为 T0 时刻, 我们记 IP_i 在 $[T0-\Delta T, T0]$ 时间段内访问过的所有域名集合为伴随域名, 那么黑产门户(图中蓝色网站)则可能出现在这一集合中. 为排除用户访问的偶然性、降低筛选伴随域名数量级, 考虑与所有访问 IP 的产生时序关联的伴随域名, 若某域名在多个 IP 的访问序列中同时出现, 那么其引流的概率也就越大.

2.2 总体流程

基于 2.1 节的设计思路, 本文提出一种基于突变流量分析的黑产应用发现方法, 该方法由 3 部分组成, 整体流程图如图 4.

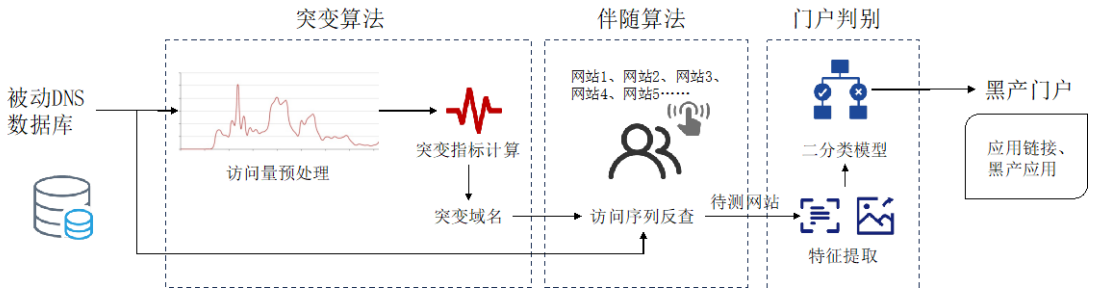


图 4 系统流程图(彩印)

首先，从被动 DNS 数据库出发，利用域名的访问量特点完成初筛，再基于突变指标计算得到突变域名. 由于互联网空间中域名数量众多，我们在计算突变前将采用域名当日及过去某段时间内解析量、访问量阈值过滤解析量和访问量较低的域名. 同时，我们设计了突变指标的计算规则，将短期访问量指标和长期访问量指标作为基准值，测算当日访问量的突变率，进而得到最终的突变域名集合. 使用访问量预处理可以有效降低筛选流程数据量，减少后续流程中不必要的计算量，提高整体运行效率.

其次，通过对突变域名的访问序列反查，建立伴随域名集合. 此步骤是本方法的关键部分，它利用了用户访问黑产门户网站和导流的黑产域名的序列在被动 DNS 数据中呈现的关联模式特征，从突变域名扩展发现门户域名. 该方法可从访问量发生突变的黑产网站关联分析至不具备明显访问量变化特征的黑产门户网站，可用于捕获黑产门户这种持续稳定的分发渠道，进而更高效地获取到黑产应用. 同时，由于被动 DNS 数据只包含了用户请求域名解析信息，不带有具体某个网站的跳转链路，因此备选集合中含有大量噪音域名，此过程需要通过相关算法控制待测网站集合规模.

最后，利用分类模型判别伴随域名中的黑产门户网站，并获取其中分发的应用. 此过程首先将针对性地对上一步骤计算出的伴随域名(即待测网站)进行多维特征提取. 在特征选取时，充分考虑了黑产门户的传播特点、视觉特点，结合图像识别技术和外链特征构造，并使用了多种机器学习模型进行训练.

2.3 突变算法描述

在本文工作中，突变概念选取本文 2.1 节的定义. 本文使用被动 DNS 数据库作为数据源，首先利用访问量预处理完成域名初筛，过滤解析量和访问量较低、被引流概率较小的域名，然后，基于突变指标计算值测算突变率，使用突变阈值 $Threshold_{short\_spike}$ 和 $Threshold_{long\_spike}$ 评估域名的突变率是否满足本文的突变概念，最终输出突变域名集合.

在域名初筛阶段，本文依据表 1 标准过滤出可能被引流的域名.

表 1 突变域名初筛规则

规则名称	规则描述	初筛方法
域名有效性	判断是否为反向域名	忽略以".in-addr.arpa"等为后缀的反向域名
域名解析	判断域名是否被解析过	访问客户端数>=1
域名泛解析	判断域名是否为泛解析域名	解析过的子域名数量<=50,000,000
域名解析量	判断域名是否有足够解析	30,000<=解析量<=1,000,000,000
域名流行度	判断是否为信誉较高的域名	Tranco[18]域名排名>1000

对完成过滤后的域名，本文定义最近两日的总请求量比率、访问客户端 IP 数比率，来反映当日访问较

前一日的突变程度:

$$Short\_Spike\_Rate_{域名总请求量}=\frac{域名请求量_{today}}{域名总请求量_{last\_day}} \tag{2}$$

$$Short\_Spike\_Rate_{访问客户端IP数}=\frac{访问客户端IP数_{today}}{访问客户端IP数_{last\_day}} \tag{3}$$

同时, 为了降低偶然性, 使用一段时间内总请求量、访问客户端 IP 总量的平均值代表该域名在被引流前的平均情况, 重定义上述变量来反映当前访问情况较过去一段时间内的突变程度:

$$Long\_Spike\_Rate_{域名总请求量}=\frac{域名总请求量_{today}}{平均每日域名总请求量_{last\_10\_days}} \tag{4}$$

$$Long\_Spike\_Rate_{访问客户端IP数}=\frac{访问客户端IP数_{today}}{平均每日访问客户端IP数_{last\_10\_days}} \tag{5}$$

本节, 我们使用 $Threshold_{short\_spike}$ 和 $Threshold_{long\_spike}$ 作为突变阈值, 同时满足以下条件的即为突变域名.

$$Short\_Spike\_Rate_{域名总请求量} > Threshold_{short\_spike} \tag{6}$$

$$Long\_Spike\_Rate_{域名总请求量} > Threshold_{long\_spike} \tag{7}$$

$$Short\_Spike\_Rate_{访问客户端IP数} > Threshold_{short\_spike} \tag{8}$$

$$Long\_Spike\_Rate_{访问客户端IP数} > Threshold_{long\_spike} \tag{9}$$

2.4 伴随算法描述

在本文工作中, 伴随概念选取本文 2.1 节的定义. 本节算法使用上一节输出的突变域名集合作为输入, 同时利用被动 DNS 作为数据源, 利用主动反查用户访问序列的方式从突变域名的时序流量信息中提取对应的伴随域名, 最终输出伴随域名集合.

获取某日突变域名的伴随域名的算法阐述如下:

(1) 输入突变域名集合 *spike_domain_set*, 伴随时间窗口长度 Δt, 伴随域名抽取阈值 $Threshold_{IP}$;

(2) 置伴随集合 *accompany_domain_set* 为空;

(3) 对 *spike_domain_set* 中的每个突变域名 $domain_{spike}$:

a) 获取在发生访问量突变当天访问过该域名的所有客户端 IP 地址 *visit_IPs*;

b) 对每个访问IP_i, 获取该客户端 IP 在时间窗口Δt 内的被动 DNS 请求序列 $Query\_Seq_i$,提取请求序列中的请求域名 $Query\_Domains_i$;

c) 对 $Query\_Domains_i$ 中的每个请求域名 $Query\_Domains_{i,j}$ 计数,计数结果存入 *Domain_Counter*;

d) *Domain_Counter* 中为所有关联 IP 的请求域名计数结果, 若某一请求域名 $domain_k$ 同时满足

$$domain_k \in visit\_IPs \tag{10}$$

$$\frac{Domain\_Counter[domain_k]}{num(visit\_IPs)} > Threshold_{IP} \tag{11}$$

说明有超过 $Threshold_{IP}$ 比例的客户端 IP 在访问突变域名 $domain_{spike}$ 前均访问了 $domain_k$, 因此 $domain_k$ 可能为造成突变域名 $domain_{spike}$ 的黑产门户域名, 则将 $domain_k$ 加入伴随集合 *accompany_domain_set*;

(4) 最终输出 *accompany_domain_set* 即为当日的所有突变域名对应的伴随域名集合.

2.5 黑产门户网站特征表示与判别

为了从大量伴随域名判别出黑产门户网站, 从黑产门户网站的引流功能出发, 深入解析引流网站的四个方面的特征:

(1) 域名元数据: 黑产门户网站通常具有稳定且较高的访问流量, 以便支撑黑产门户为其他网站导流, 这也是黑产门户能够通过放置推广链接盈利的原因. 表现在特征上, 黑产门户通常具有较高的访问量绝对值和较小的方差. 此外, 网站的访问总量和访问 IP 数分别反映了网站被请求的总次数和访问的用户数量, 在实践中可以代表网站流量和实际影响人数, 因此本文同时分析这两个量;

(2) 网页外链特征: 黑产门户网站往往含有比普通网站更多的外链数、渠道数. 外链是指此网站之外的外部链接, 外链数越多, 其引流特征越明显. 渠道是一类特殊的用于计算网站访问量的外链,通常由第三方数据统计平台(如百度统计[19])提供服务, 由渠道统计得到的访问量往往作为结算门户网站推广费的依据,为此,黑产门户中嵌入的渠道量往往高于普通网页. 由于黑产门户分发平台的特性, 其内容外链往往来自不同的二级域名(Second-Level Domain, SLD)、不同的完全限定域名(Fully Qualified Domain Name, FQDN), 当 FQDN 相同的时候, 外链路径(path)不同时也会传播不同的黑产应用, 因此以上特征均能表现黑产门户分发移动应用的能力.

(3) 网页源码特征: 网页布局、网页主题等信息可以特异性指征黑产门户. 首先, 由于黑产门户的设计初衷是用于大量展示并分发应用, 因此在页面结构上倾向于频繁使用列表等并列结构. 其次, 黑产门户不同于普通门户, 其页面内容中具有明显的黑产相关语义, 以及诱导用户下载、点击的引导语义, 其中文本语义特点可供表征其黑产特性.

(4) 网页截图特征: 黑产门户网站需要吸引用户点击以增加访问量, 这一需求反映到网页图像上则呈现出明亮、鲜艳等特征. 本方法在爬取待测网站的同时保存了网页截图, 利用截图中的视觉效果, 将图像转化为可被量化的指标, 基于不同算法提取图像的亮度、饱和度、对比度可用于区分黑产门户.

最终本文在以上四个方面共提取到了 21 个特征作为黑产门户的识别依据, 具体特征及特征类型见表 2.

表 2 黑产门户识别特征及特征类型

特征类别	特征内容	特征类型
域名元数据	30 天内访问量	数值序列
	30 天内访问量方差	数值
	30 天内访问 IP 数	数值序列
	30 天内访问 IP 数方差	数值
网页外链内容	外链数量	数值
	外链中图片的比例	数值
	外链中的 SLD 数量	数值
	SLD 数量/总外链数量	数值
	外链中的 FQDN 数量	数值
	FQDN 数量/SLD 数量	数值
	外链的 path 数/FQDN 数量	数值
	渠道数	数值

网页源码内容	列表数	数值
	最大列表宽度	数值
	“下载”相关语义词数	数值
	“黑产”相关语义词数	数值
网页截图	亮度	数值
	亮度(luminosity 方法)	数值
	饱和度	数值
	空白对比度	数值
	对比度	数值

在提取到上述特征以后, 本方法将利用机器学习模型对提取到的特征进行处理, 精确识别黑产门户网站, 并输出识别结果. 为避免不同维度的数据分布、数量级、单位的不兼容性, 因此选择采用不依赖量纲的随机森林分类模型. 由于决策树的简洁性和森林构建的随机性, 该算法相比其他算法能够有效地在大数据集上运行, 并且可以判断不同特征的重要程度和相互的影响, 泛化能力强. 对于不同类型机器学习模型的效果评估见 3.2 节 RQ2. 具体而言, 本文首先构建了多个互不关联的二分类决策树, 将它们通过随机方式建立成为森林, 在每次训练中, 利用 Bagging 策略算法从汇总得到的各分类树结果中计算并输出精确的甄别结果. 通过对不同决策树数量及最小叶节点样本数、最小分割叶节点样本数进行了梯度实验, 确定最终的随机森林模型中最小叶节点样本数为 1, 最小分割叶节点所需样本数为 8, 决策树数量为 200.

3 实验分析

本节以 2023 年 5 月 1 日至 2023 年 5 月 31 日共 31 天的被动 DNS 流量为数据源开展研究, 通过一系列实验, 评估本文提出的基于突变流量分析的黑产应用发现方法的有效性. 本节首先介绍相关实验方法和主要评估指标(第 3.1-3.2 节), 然后围绕以下 5 个研究问题开展实验验证(第 3.3 节).

RQ1: 本文方法累计发现的黑产门户及应用有多少?

RQ2: 本文方法效率相比基准方法有多大提升?

RQ3: 本文方法是否可以有效检测黑产门户?

RQ4: 本文方法是否可以有效获得零天应用?

RQ5: 本文方法捕捉到的应用类型有哪些?

3.1 实验方法

本实验使用来自某匿名的安全公司提供的被动 DNS 数据源(包含中国最大的 DNS 提供商 114DNS 的公共 PDNS 数据), 实验时间为 2023 年 6 月 15 日-24 日共 10 天, 运行服务器环境为 5 台 Ubuntu 18.10 服务器(16 核,32G 内存)执行分布式网络爬虫和模型判别的任务, 和 1 台 Ubuntu 18.04 服务器(32 核,224G 内存)执行流水化任务管理和数据分析的管理.

对于突变算法模块和伴随算法模块, 本实验选取 2023 年 5 月 1 日至 2023 年 5 月 31 日共 31 天的被动 DNS 流量数据作为突变域名的数据源, 应用本方法自动化计算突变、伴随域名. 为平衡计算和网络资源限制, 保证实验顺利进行, 我们设置伴随时间窗口长度Δt为 120 秒, 伴随域名抽取阈值$Threshold_{IP}$为 30%, 最终平均每日突变伴随的计算时间为 7.74 小时.

对于门户判别模块, 我们在前期研究中人工标注了 500 个黑产门户网站样本, 选取 Tranco[18]中前 500 个域

名作为网站白样本. 在模型训练阶段, 我们将数据集按照 4:1 划分为训练集和测试集, 并使用五折交叉验证的方法评估模型效果. 在实际运行阶段, 我们选取效果最佳的分类器模型, 并对域名进行特征提取, 最后判别是否为门户域名.

为串联本实验多个流程, 提高运行效率, 我们采用了 redis 作为流水线任务管理工具, 使用 Mysql、SeaweedFS 作为文件存储数据库, 设置 batch_size 为 100, 每台任务服务器从 redis 中读取 100 个任务后分别完成对应任务(如爬虫任务、特征提取任务或模型判别任务), 并将数据结果写入数据库内, 多任务的批量执行减少了每个任务的网络往返时间和服务器处理时间, 从而提高了任务处理的整体性能.

3.2 实验评价指标

本实验中在进行黑产门户网站判别时涉及机器学习模型训练. 该过程为二分类任务. 我们采用准确率、精确率、召回率、F1 值作为判别结果评估指标. 计算方法如下:

$$准确率=\frac{TP+TN}{TP+TN+FP+FN},精确率=\frac{TP}{TP+FP} \tag{12,13}$$

$$召回率=\frac{TP}{TP+FN},F1值=\frac{2TP}{2TP+FP+FN} \tag{14,15}$$

3.3 实验结果与分析

• RQ1:本方法累计发现的黑产门户及应用有多少?

本方法运行结果如表 3 所示, 此表展示了在系统测试的 31 日中, 每日发现突变域名、伴随域名、黑产门户、下载应用等数量的平均值、最大值和最小值. 结果显示, 从每日被动 DNS 数据源出发, 经过突变伴随算法、黑产门户判别后, 平均每日能发现 2222.48 个黑产门户、下载 1079.19 个应用, 平均每个门户能下载到 0.48 个应用. 考虑不同门户可能会分发相同应用, 将收集到的应用按 MD5 去重后, 平均每日可以获取到 751.48 个不同的应用. 本方法在 31 天内一共识别出 10,921 个黑产门户, 获取 3,439 个不同的应用下载链接和 3,303 个不同的移动应用.

进一步地, 本文还分析了捕获应用与黑产门户对应关系. 测试中, 我们观测到分发应用数最多的一个门户累计分发了 55 个不同的应用, 该现象表明当前黑产应用通过门户进行集中传播.

表 3 发现黑产门户及应用数量(单位:个)

统计量	突变域名	伴随域名	黑产门户	捕获应用	捕获应用(去重)
平均值	8746.96	121803.64	2222.48	1079.19	751.48
最小值	6237	87833	1759	635	475
最大值	11122	147151	2608	2550	1381

• RQ2: 本文方法效率相比基准方法有多大提升?

为了对本文所提出的应用采集方法的效率进行评估, 本文采用通常使用的基于搜索引擎的黑产应用采集技术作为基准方法. 具体而言, 本次测试中我们选取在中文互联网环境中常用的百度、搜狗、360 搜索引擎进行实验. 对于每个搜索引擎, 我们采用 50 个黑产相关的关键词与“应用下载”结合, 如“博彩 应用下载”、“娱乐城 应用下载”等, 并对每组关键词选取前 50 个搜索结果. 之后使用 RQ1 中相同的流程爬取网站内的所有应用, 其结果如表 4 所示. 上述三家搜索引擎每家爬取了 2500 个网站, 平均采集到 21 个应用, 其中最多的是百度, 爬取到了 40 个. 基准方法的平均应用采集率(捕获应用数/爬取网站数)为 0.0084, 远低于本文方法的应用采集率 0.4856.由此表明, 采用基于突变流量分析的方法可以明显提升应用采集率 57.8 倍.

表 4 基于搜索引擎的黑产应用采集结果对比(单位:个)

采集渠道	疑似网站	捕获应用	应用采集率
百度搜索引擎	2500	40	0.016
搜狗搜索引擎	2500	20	0.008
360 搜索引擎	2500	3	0.0012
搜索引擎(平均)	2500	21	0.0084
本文方法(平均)	2222.48	**1079.19**	**0.4856**

• RQ3:本文方法是否可以有效检测黑产门户?

为了更加准确地判别黑产门户, 本文测试了多种不同的机器学习模型算法, 并对每种算法的参数进行优化, 最终筛选出效果最佳的分类器模型算法及其对应的最优参数.

具体而言, 我们选取了 6 种较为经典的机器学习分类模型, 分别是决策树模型[20],随机森林模型[21], KNN 模型[22],SVC 模型[23],ADA 模型[24]和 MLP 模型[25], 并将准确率、精确率、召回率、和 F1 值作为指标评估模型的效果, 以上指标已在 3.2 节中定义. 通过调整每个模型的关键参数, 最终得到的测试结果见表 5 至表 10.

表 5 决策树模型测试结果

(h 为最大高度, n 为最小叶子节点样本数, s 为最小叶子节点分隔数)

(h,n,s)	准确率	精确率	召回率	F1
(3, 1, 3)	0.945	0.989	0.900	0.942
(3, 1, 5)	0.945	0.989	0.900	0.942
(3, 3, 3)	0.945	0.989	0.900	0.942
(3, 3, 5)	0.945	0.989	0.900	0.942
(5, 1, 3)	0.945	0.968	0.920	0.944
(5, 1, 5)	0.955	0.979	0.930	0.954
(5, 3, 3)	0.955	0.979	0.930	0.954
(5, 3, 5)	0.955	0.979	0.930	0.954

表 6 KNN 模型测试结果

(k 为最近邻居的数量, n 为叶子节点的大小)

(k,n)	准确率	精确率	召回率	F1
(3, 20)	0.785	0.766	0.820	0.792
(3, 30)	0.785	0.766	0.820	0.792
(3, 40)	0.785	0.766	0.820	0.792
(5, 20)	0.730	0.730	0.730	0.730
(5, 30)	0.730	0.730	0.730	0.730
(5, 40)	0.730	0.730	0.730	0.730
(7, 20)	0.735	0.742	0.720	0.731
(7, 30)	0.735	0.742	0.720	0.731

表 7 随机森林模型测试结果

(n 为最小叶子节点样本数, s 为最小分隔样本数, m 为决策树数量)

(n,s,m)	准确率	精确率	召回率	F1
(1, 4, 100)	0.955	0.950	0.960	0.955
(1, 4, 200)	0.960	0.951	0.970	0.960
(1, 8, 100)	0.955	0.950	0.960	0.955
(1, 8, 200)	**0.970**	**0.970**	**0.970**	**0.970**
(2, 4, 100)	0.955	0.950	0.960	0.955
(2, 4, 200)	0.960	0.960	0.960	0.960
(2, 8, 100)	0.955	0.950	0.960	0.955
(2, 8, 200)	0.955	0.950	0.960	0.955
(2, 9, 200)	0.955	0.950	0.960	0.955

表 8 ADA 模型测试结果

(M 表示弱分类器数量, α 表示权重缩减系数)

(M, α)	准确率	精确率	召回率	F1
(50, 0.01)	0.865	0.974	0.750	0.847
(50, 0.1)	0.935	0.939	0.930	0.935
(50, 1)	0.915	0.928	0.900	0.914
(100, 0.01)	0.880	0.975	0.780	0.867
(100, 0.1)	0.940	0.949	0.930	0.939
(100, 1)	0.920	0.929	0.910	0.919
(150, 0.01)	0.935	0.958	0.910	0.933
(150, 0.1)	0.945	0.949	0.940	0.945
(150, 1)	0.920	0.938	0.900	0.918

表 9 MLP 模型测试结果

(H 表示隐藏层的大小, σ 表示激活函数)

(H, σ)	准确率	精确率	召回率	F1
(50,tanh)	0.855	0.890	0.810	0.848

表 10 SVC 模型测试结果

(C 表示惩罚参数, k 表示核函数)

(C,k)	准确率	精确率	召回率	F1
(0.1,linear)	0.850	0.850	0.850	0.850

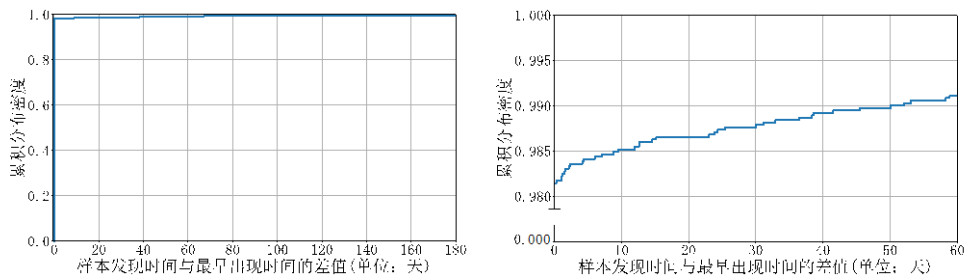
(50,relu)	0.870	0.911	0.820	0.863	(0.1,rbf)	0.675	0.664	0.710	0.686
(100,tanh)	0.870	0.911	0.820	0.863	(1,linear)	0.900	0.935	0.860	0.896
(100,relu)	0.855	0.918	0.780	0.843	(1,rbf)	0.755	0.726	0.820	0.770
150,tanh)	0.860	0.900	0.810	0.853	(10,linear)	0.940	0.940	0.940	0.940
(150,relu)	0.860	0.900	0.810	0.853	(10,rbf)	0.865	0.869	0.860	0.864

通过对比模型测试结果，我们发现，随机森林模型的效果显著优于其他模型算法，在选取(1,8,200)作为参数的情况下，该模型的准确率和精确率高达 97%，能够准确地判别门户域名.

• RQ4:本文方法是否可以有效获得零天应用?

面对执法部门和安全公司的打击和检测，黑产应用的更新十分迅速. 为此，作为黑产应用的采集方法，本研究是否能够采集到鲜活的样本尤为重要. 为科学地评估本文方法是否能够有效捕获零天应用，我们将全球最大的恶意软件共享平台 VirusTotal[13]中记录到的该应用首次上传时间近似为样本最早出现的时间，并将该时间与本方法采集到的样本发现时间进行对比，统计差值，得到了相应的累计密度分布图.

图 5-(a)记录了差值分布的整体趋势，可以看到，我们下载的应用中，有 98.14%的样本之前从未被 VirusTotal 平台采集，该结果表明我们的方法能够有效地获取零天应用. 进一步地，为了能够更清晰地观察差值的累计概率变化趋势，我们挑选了差值在 60 天内的分布结果进行分析，如图 5-(b)所示. 从图中我们可以发现，本方法捕获的应用中仅有约 1%的样本为非零天应用，平缓的增加趋势表明这些应用的上传日期存在一定的随机性，从侧面反映了本文的方法能够更加集中、高效地发现零天应用.



(a) 180天分布　　(b) 60天分布

图 5 采集应用累计密度分布(彩印)

• RQ5:本文方法捕捉到的应用类型有哪些?

前文实验已经证明，利用本方法能够有效捕捉到大量零天在野应用. 为了进一步论述本采集方法的有效性，本实验将评估利用本文方法所采集到的应用的类型. 特别地， 我们选取了当前综合了大量病毒软件检测引擎结果的 VirusTotal 作为判断应用类型的检测平台.

具体而言，我们将采集到的应用上传到 VirusTotal 平台，并通过解析报告内容，可以得到 VirusTotal 中集成的 75 个反病毒引擎对该应用的检测结果，并从中提取应用所对应的恶意标签，例如“virus”,“trojan”,“scam app”等.

统计发现，在上传的 3303 个应用中，有 3026 个应用被平台判定为恶意应用，其中部分应用甚至被检测出多种不同类型的恶意行为，例如 MD5 值为 6624151114ba45628c7acd59fa2700c2 的应用检测出的报告中含有“Virus”、“Trojan”和“PUA”等多种恶意标签. 为了进一步了解应用的恶意标签分布情况， 我们筛选出识别恶意应用数量排名前五的反病毒引擎，并统计其提供的恶意标签，得到了如图 6 所示的恶意标签分布图. 我们观察到， IKARUS 检测到的恶意应用最多，占比 58.2%. 从整体结果上看，标签为 PUA(46.9%)、

Trojan(31.8%)和 Boogr(28.4%)的样本数量最多. 其次, 有不少样本被标识为 Adlibrary(24.3%)、Riskware(19.3%)软件, 表明此类应用有一定的潜在风险. 除此之外, 还有部分样本被判别为 Adware(5.2%)、Dropper(0.4%)等其他类型的恶意应用. 此外, 由于各反病毒引擎对于样本的恶意标签并未存在统一标准, 因此图中所示的样本标签可能存在不同, 但该不一致情况并不影响本文方法检测在野黑产样本的有效性.

总体而言, 我们从门户中下载到的样本绝大多数(91.6%)都属于恶意应用, 证明了本文提出的方法在黑产应用采集方面的有效性.

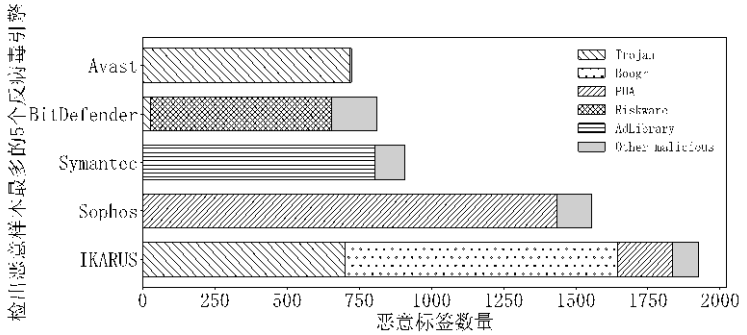


图 6 采集样本恶意标签分布图

4 总 结

在本研究中, 我们对当前网络黑产应用的分发传播现象开展了深入探讨, 并提出了一种有效的大规模采集在野黑产应用的方法, 旨在解决当前研究中移动应用数据集收集困难的问题. 根据网络黑产应用分发过程中的特点, 本文提出的基于突变流量分析批量捕获黑产应用的方法, 利用黑产门户这一传播渠道中的关键环节, 显著改善了在野黑产应用采集过程的效率和质量, 为后续的实时分析与追踪提供了强有力的数据支持.

我们的测试结果验证了该方法的有效性, 成功获取了 3,303 个不同的应用. 被捕获的移动应用中, 91.61%的样本被至少一个引擎检测为恶意软件, 而 98.14%的样本为首次采集发现的零天应用, 这为后续的黑产应用的研究和分析提供了珍贵的数据资源.

References:

[1] 360 数字安全. 《2022 年度反诈报告》重磅发布! 2023. https://wlaq.gmw.cn/2023-02/18/content_36375612.htm

[2] 中国政府网. 公安部组织开展新一轮集中收网行动依法严厉打击涉电信网络诈骗 APP 技术开发违法犯罪团伙 [EB/OL]. 2021. http://www.gov.cn/xinwen/2021-05/12/content_5605957.htm

[3] Yang H, Du K, Zhang Y, et al. Casino royale: a deep exploration of illegal online gambling[C]//Proceedings of the 35th Annual Computer Security Applications Conference. 2019: 500-513.

[4] Lee P Y, Hui S C, Fong A C M. An intelligent categorization engine for bilingual web content filtering[J]. IEEE Transactions on multimedia, 2005, 7(6): 1183-1190.

[5] Ali F, Khan P, Riaz K, et al. A fuzzy ontology and SVM–based Web content classification system[J]. IEEE Access, 2017, 5: 25781-25797.

[6] Platzer C, Stuetz M, Lindorfer M. Skin sheriff: a machine learning solution for detecting explicit images[C]//Proceedings of the 2nd international workshop on Security and forensics in communication systems. 2014: 45-56.

[7] Wehrmann J, Simões G S, Barros R C, et al. Adult content detection in videos with convolutional and recurrent neural networks[J]. Neurocomputing, 2018, 272: 432-438.

[8] Perez M, Avila S, Moreira D, et al. Video pornography detection through deep learning techniques and motion informati on[J]. Neurocomputing, 2017, 230: 279-293.

[9] Shodan. Shodan: Search Engine for the Internet of Everything. https://www.shodan.io/

[10] Knownsec. Zoomeye – Cyberspace Search Engine. https://www.zoomeye.org/

[11] Gao Y, Wang H, Li L, et al. Demystifying illegal mobile gambling apps[C]//Proceedings of the Web Conference 2021. 2021: 1447-1458.

[12] Hong G, Yang Z, Yang S, et al. Analyzing Ground-Truth Data of Mobile Gambling Scams[C]//2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022: 2176-2193

[13] VirusTotal. VirusTotal - Free Online Virus, Malware and URL Scanner. https://www.virustotal.com/

[14] Milly F. What is Passive DNS? A beginner's guide. 2022. https://www.spamhaus.com/resource-center/what-is-passive-dns-a-beginners-guide/

[15] Reverse IP/DNS Blog & How To Guides to Obtain Reverse IP/DNS Data. 2022. https://dns-history.whoisxmlapi.com/blog/passive-dns.

[16] 中国国家互联网信息办公室.“网络黑产”到底是啥？普通人应怎样防范？这些知识你要知道. 2023. http://www.cac.gov.cn/2019-09/17/c_1570248615898997.htm

[17] OF INVESTIGATION F B. 2020 internet crime report[EB/OL]. 2021. https: //www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf.

[18] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019). https://doi.org/10.14722/ndss.2019.23386.

[19] 百度. 网站分析白皮书. https://tongji.baidu.com/web5/image/%E7%99%BE%E5%BA%A6%E5%8F%91%E5%B8%83%E3%80%8A%E7%BD%91%E7%AB%99%E5%88%86%E6%9E%90%E7%99%BD%E7%9A%AE%E4%B9%A6V3.0%E3%80%8B.pdf

[20] Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.

[21] Biau G, Scornet E. A random forest guided tour[J]. Test, 2016, 25: 197-227.

[22] Peterson L E. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.

[23] Gunn S R. Support vector machines for classification and regression[J]. ISIS technical report, 1998, 14(1): 5-16.

[24] Schapire R E. Explaining adaboost[M]//Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 37-52.

[25] Gardner M W, Dorling S R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences[J]. Atmospheric environment, 1998, 32(14-15): 2627-2636.

附中文参考文献:

[1] 360 数字安全. 《2022 年度反诈报告》重磅发布！2023. https://wlaq.gmw.cn/2023-02/18/content_36375612.htm

[2] 中国政府网. 公安部组织开展新一轮集中收网行动依法严厉打击涉电信网络诈骗 APP 技术开发违法犯罪团伙 [EB/OL]. 2021. http://www.gov.cn/xinwen/2021-05/12/content_5605957.htm

[16] 中国国家互联网信息办公室.“网络黑产”到底是啥？普通人应怎样防范？这些知识你要知道. 2023. http://www.cac.gov.cn/2019-09/17/c_1570248615898997.htm

[19] 百度. 网站分析白皮书. https://tongji.baidu.com/web5/image/%E7%99%BE%E5%BA%A6%E5%8F%91%E5%B8%83%E3%80%8A%E7%BD%91%E7%AB%99%E5%88%86%E6%9E%90%E7%99%BD%E7%9A%AE%E4%B9%A6V3.0%E3%80%8B.pdf