SOTI

第 10 卷，第 3 期

10 YEARS
OF SECURITY INSIGHT

侵蚀您的利润

网络爬虫程序对电商行业有何影响

Akamai

互联网现状/安全性

目录

您知道吗？超过一半的网络流量都来自于爬虫程序。特别是商业垂直行业，由于该行业依赖 Web 应用程序和资产创收，因此他们受高风险爬虫程序流量的影响最大（图 1）。尽管我们常常听到爬虫程序在不断进化，但当前的电子商务类企业特别关注的是网络抓取类爬虫程序，因为它们的经济影响往往隐藏在表面之下，与其他类型的爬虫程序截然不同。随着人工智能 (AI) 僵尸网络和无界面浏览器技术的崛起，抓取类爬虫程序越来越难以检测，也就变得极其难以躲避。举个例子，Akamai 的一家电商客户在不知不觉中拦截的高达 99% 的高风险流量就源自于抓取类爬虫程序。

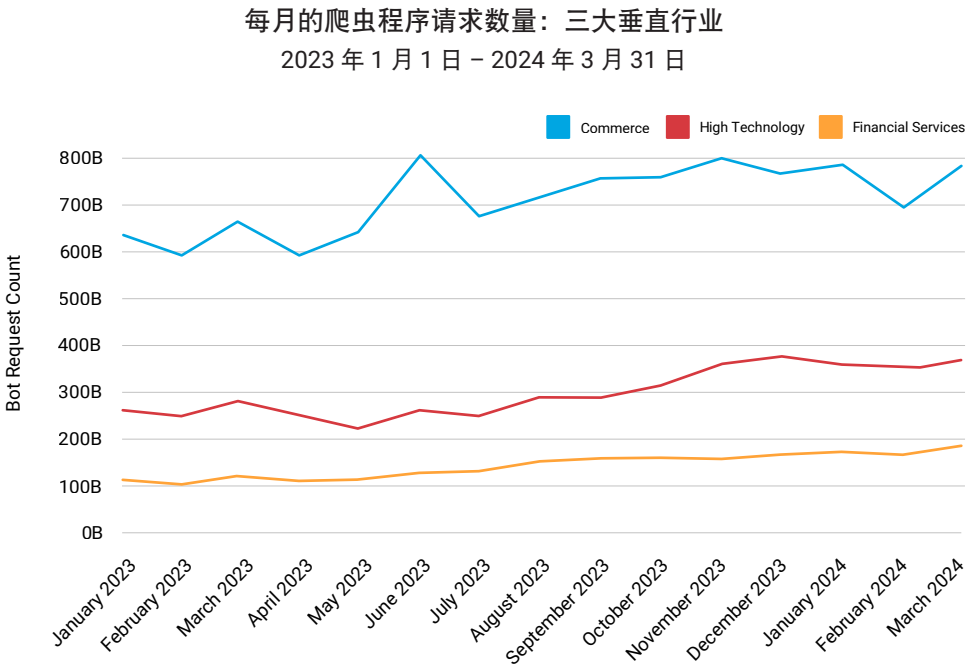

图 1：商业是收到爬虫程序请求数量最多的垂直行业，从 2023 年初至 2024 年第一季度，全球商业垂直行业的爬虫程序流量呈现出上升趋势

因此，在这份《互联网现状》(SOTI) 报告中，我们将聚焦于这些爬虫程序及其操纵者的演变与专业化趋势。尽管爬虫程序的存在由来已久，但我们依然观察到各种非法团伙在利用这种手段来实施犯罪攻击、欺诈活动以及窃取竞争情报。最近，我们观察到爬虫程序的使用呈现普遍增长的趋势，其中抓取类爬虫程序对企业造成的负面影响正在加剧。本报告旨在提供技术见解和应对策略，以加深整个商业行业对这一日益严峻问题的认识。

爬虫程序：良性、恶意和中性

每家以电商为主的大型企业都面临着爬虫程序带来的挑战，这些爬虫程序不断演变进化，变得更加专业化，以实现其攻击目的。在商业垂直行业中，存在着各种各样的爬虫程序，它们各自承担着不同的任务。为了方便理解，可以将它们分为三类：良性爬虫程序、恶意爬虫程序和中性爬虫程序。良性爬虫程序有助于客户找到您的网站。恶意爬虫程序则出于恶意目的从您的网站抓取内容。中性爬虫程序往往比较活跃，但仍然是合法的。它们实际上是良性爬虫程序的一个子类（比如频繁发送 ping 回显请求的合作伙伴爬虫程序或频繁发出调用请求的程序 API）。

考虑到聊天机器人和搜索引擎类爬虫程序所带来的诸多益处，比如回答用户的基本问题和提供能够返回更准确搜索结果的网站内容，我们希望在控制 IT 成本的同时，能够优化这类爬虫程序的性能。对于那些恶意爬虫程序，比如未经授权尝试访问客户帐户以进行帐户接管的撞库爬虫程序，我们希望在不损害整体客户体验的前提下采取防范措施。近期，网络抓取类爬虫程序已成为亟待解决的一大难题，它们会导致收入减少、忠诚度下滑，以及成本不断攀升。

抓取类爬虫程序是一种用于直接从互联网的网站上提取数据和内容的僵尸网络，它非常特别。网络抓取类爬虫程序引起了人们的高度关注，因为它们的运作方式、对业务的影响以及检测难度都有别于其他爬虫程序。网络抓取类爬虫程序的应用场景也多种多样，具体取决于企业和爬虫程序操纵者如何利用这些爬虫程序收集的信息来创收。不论具体的目标为何，抓取类爬虫程序都会导致收入减少、IT 成本攀升以及整体客户体验降低。

在这份 SOTI 报告中，我们探讨了内容抓取对电商行业的影响，并分析了为何相关的业务负责人（如数字、营销、品牌、财务、风险和安全等领域）应共同关注并采取措施来遏制滥用的抓取类爬虫程序。为了更好地摸清这些影响，我们需要深入了解网络抓取类爬虫程序的演变动因、使用目的、运行机制及其所带来的影响，同时探讨商企可以采取哪些措施来应对这一问题。

报告的关键见解

网络内容抓取不仅仅关乎欺诈或安全问题，它还是一个商业问题。抓取类爬虫程序会对企业的多个关键方面产生负面影响，包括减少收入、削弱竞争优势、损害品牌形象、降低客户体验、增加基础架构成本以及破坏数字化体验，不一而足。

根据 Akamai 的案例研究，我们发现 42.1% 的流量活动来自爬虫程序，而在这些爬虫程序流量中，高达 65.3% 的流量来自恶意爬虫程序。总计有 63.1% 的恶意爬虫程序流量运用了先进的技术。

无界面浏览器技术改变了抓取类爬虫程序的现状，为了管理这种爬虫程序活动，我们需要一种比其他基于 JavaScript 的抵御措施更为精细的策略。

无论抓取类爬虫程序是出于恶意还是善意目的抓取企业数据，都可能导致企业遭受包括网站性能下降、网站指标污染、网络钓鱼网站发起的盗用凭据攻击，以及计算成本攀升等技术影响。

观察和了解不同的流量模式至关重要，这有助于我们判断一个网站是吸引了人类访客、基础爬虫程序还是复杂的爬虫程序流量。这些模式包括昼夜波动、间歇性和连续不断。

良性爬虫程序与恶意爬虫程序的对比

我们首先了解基础知识：爬虫程序，也称为“机器人”，它是一种计算机程序，能够比人类更迅速、更准确地执行自动化任务。爬虫程序根据其角色和类型，主要可以分为两大类：良性爬虫程序和恶意爬虫程序（图 2）。为了简化这种比较，我们将中性爬虫程序与良性爬虫程序合并，因为中性爬虫程序实际上是良性爬虫程序的一个子类。

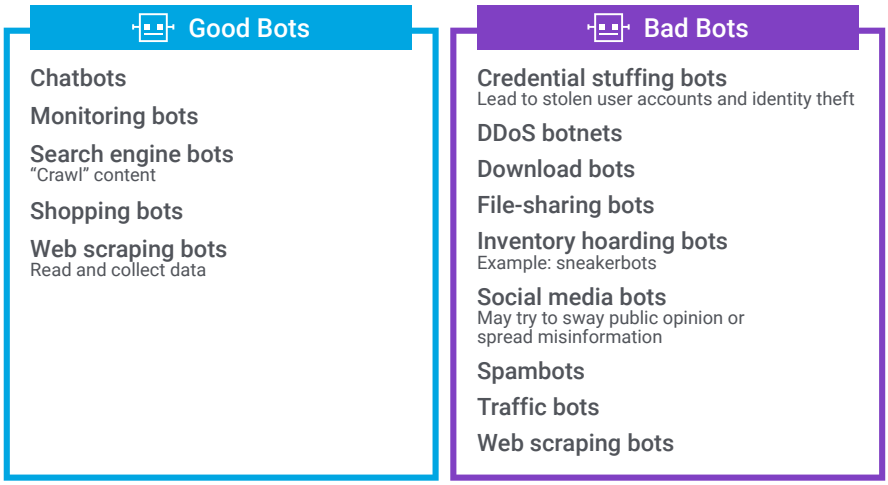

图 2：良性爬虫程序和恶意爬虫程序的并排比较以及示例

良性爬虫程序是有助于提供工具和服务的实用爬虫程序。而与之相对的，恶意爬虫程序是指被网络犯罪分子和欺诈者恶意利用的爬虫程序。流量爬虫程序就是一个典型的恶意爬虫程序示例，它通过模拟人类在线行为来增加网站的点击量和流量（例如，实施广告欺诈）。

网络内容抓取爬虫程序有善有恶。关键在于企业如何运用这些爬虫程序收集的信息。接下来，我们将更深入地探讨全球一些大型零售商和电子商务品牌所面临的与抓取类爬虫程序相关的各种应用场景，包括其带来的正面和负面影响。

抓取类爬虫程序的基本概念

网络内容抓取是电商企业广泛使用的一种技术。以旅游和酒店行业为例，旅游聚合商从酒店和航空公司合作伙伴的网站上抓取动态内容，从而获取最新的供应情况和价格信息。这种抓取行为是有利的，当真实的用户想要进行预订操作时，企业会采取一系列常见的爬虫程序控制措施，以限制抓取类爬虫程序。企业还会借助数据提取服务提供商，从竞争对手那里收集潜在客户和其他相关的信息。此外，抓取类爬虫程序还可以用于分析数据以识别趋势。抓取类爬虫程序还有助于审查网站内容，从而改进在线产品和服务的质量，以及让潜在消费者能够通过搜索引擎等途径，更便捷地找到公司的产品。所有这些行动都能增强企业的竞争力。然而，我们也不能忽视一个事实，那就是许多实体出于不那么高尚的目的而使用了抓取类爬虫程序。

抓取类爬虫程序引起关注，客户开始警觉

令人遗憾的是，我们时常听闻消费者不慎成为网络钓鱼骗局的受害者。在这种情况中，诈骗者可能会利用抓取类爬虫程序来抓取产品图像、描述和定价信息，从而创建虚假的店面或网络钓鱼网站，企图骗取帐户凭据或信用卡信息。这些网络钓鱼/假冒网站是品牌仿冒的一种手段，它们滥用受害企业的知识产权，以此骗取潜在客户的信任。

目前，全球一些大型电子商务品牌已经遭受了假冒网站、网络钓鱼活动的侵害，也曾因在品牌仿冒活动中公司网络数据被窃取而蒙受损失（图 3）。不幸的是，当网络钓鱼网站得手时，这会导致合法品牌失去客户的信任与忠诚度。

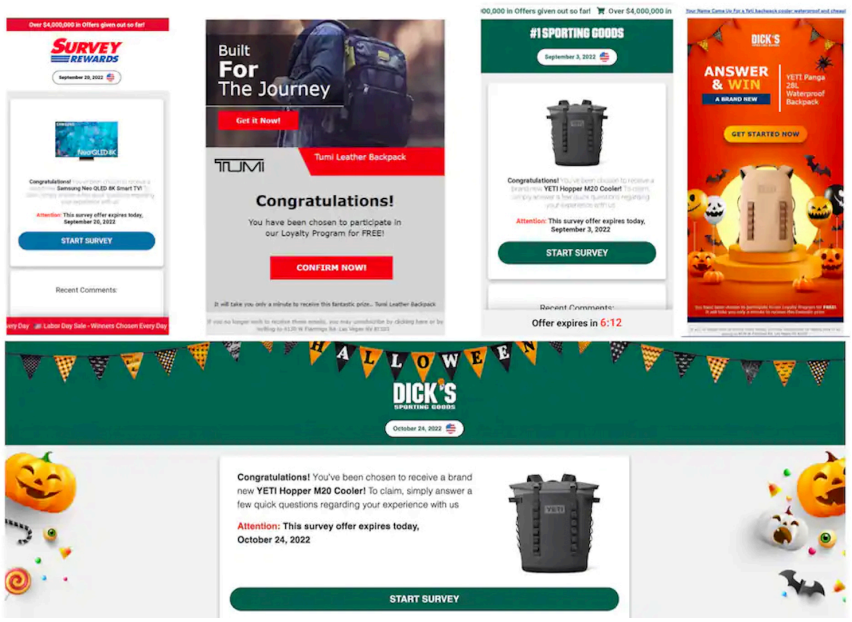

图 3：某些大型电子商务公司成为品牌仿冒活动受害者的示例

抓取类爬虫程序属于网络内容抓取技术的一种，它们能够抓取网站上销售的产品，将其抢购一空，让合法客户失去购买机会（图 4）。

图 4：抓取类爬虫程序应用场景

执行这类恶意抓取活动的攻击者非常清楚，他们的恶意行为对受害者将造成多么严重的影响。这些影响包括竞争情报/间谍活动、库存囤积/抢购、仿冒和冒牌网站/商品，以及媒体网站搜索和转贴的负面影响（表 1）。而且，目前还没有明确禁止使用抓取类爬虫程序的法律。

影响	说明
竞争情报/间谍活动	竞争对手利用企业网站上的信息来压低价格，改变自身报价，并了解新机会和威胁。
库存囤积/抢购	黄牛党不断向目标网站发出 ping 回显请求，从中寻找有货的产品，然后将其添加到购物车中，导致真正的客户买不到这些产品。
仿冒和冒牌网站/商品	仿冒者利用抓取到的内容制作虚假网站和产品目录，以此诱骗用户，导致用户误以为买到的是合法商品，而非仿冒商品。
媒体网站搜索和转贴	攻击者可以爬取新闻文章、博客和其他内容，并将其发布到自己的网站上，从而分掉原公司的访问流量和潜在的广告收益。 广告投放率通常根据网站的访客/受众量来设定，因此，当访客数量减少时，媒体网站就会错失通过原本较高的广告投放率带来的潜在收益。

表 1：网络抓取类爬虫程序故意造成的负面影响

网络内容抓取的一般附带后果

不论网络内容抓取的意图为何，企业都需要应对其副作用所带来的成本。一些公司可能会为有价值的抓取服务支付费用，然而，那些被抓取内容的公司也需要自行承担由此产生的代价。这些代价包括反爬虫解决方案费用，以及网站性能下降和关键指标被污染所带来的负面经济影响（表 2）。

影响	说明
增加服务器、CDN 和云成本，为爬虫程序流量提供服务	这不仅会影响收益，还可能因为竞争对手、攻击者以及仿冒者利用抓取到的内容，给声誉造成损害。
网站性能降级	由于抓取类爬虫程序会持续运行，一直到最终停止，所以它们会增加服务器成本和交付成本，导致企业为不需要的爬虫程序流量提供服务，并降低用户体验，例如网站响应速度缓慢、应用程序性能低下。
关键指标污染	未被发现的爬虫程序活动会严重影响网站转化率等关键指标，而业务团队需要依靠这些指标来做出有关产品定位战略和营销活动的投资决策。

表 2：网络抓取类爬虫程序无意造成的负面影响

出租抓取类爬虫程序：第三方网络内容抓取服务

正如我们之前所讲的，网络抓取类爬虫程序亦正亦邪。与那些用于撞库攻击的爬虫程序（此类爬虫程序会因具有明显的恶意企图而被阻止）不同，有些公司会提供合法的网络内容抓取服务。许多企业利用第三方网络内容抓取服务来提取和提供数据，以满足自身业务需求，这在竞争激烈的市场营销环境中是有益的。

目前，市场上有数十家提供各种网络内容抓取/数据提取服务的公司，甚至还有专门推广此类服务的研讨会。例如，Bright Data 举办了一场名为 ScrapeCon 的会议，该会议汇集了众多研究规避爬虫程序检测技术的专家，旨在帮助公司学习如何抓取数据。表 3 中包含了一些示例，展示了第三方网络内容抓取公司可能提供的不同服务级别。

服务级别 1	抓取操作中包含代理服务，这些代理服务所提供的基础架构可能包含数据中心使用的移动 IP 地址以及住宅地址。
服务级别 2	二级还涵盖自动数据提取，对数据进行净化和结构化处理，旨在使客户的数据科学团队成员能够更轻松地使用数据，进而提取有价值的情报来指导业务决策。
服务级别 3	最高级别增加了实际商业情报提取，进一步增强企业的决策过程。这些被统称为“AI 僵尸网络”。

表 3：第三方网络内容抓取公司提供的不同服务级别

客户可以自由选择从最基本到最高级的服务，并设定数据收集的频率。同时，他们还可以明确指定自己的目标。通常，所提供的服务级别或所选的僵尸网络，取决于客户需要突破的保护级别。较基础的僵尸网络能够借助位于数据中心内的数千个代理服务器运行高级脚本，从而进行数据收集并平衡流量负载。如果保护级别相对简单，则僵尸网络可以运用这项技术，穿过安全基础架构的爬虫程序管理防御措施和 Web 应用程序防火墙。

然而，当保护级别较高时，则可能需要采取更为精密的抓取策略，如无界面浏览器攻击。不论行为者执行抓取活动是出于善意还是恶意，这一点都同样适用。此外，这并非一项经济实惠的选择，因为公司需要为更复杂的基础架构支付远高于基本服务级别的成本。高级防御可能包含多种挑战技术（如验证码或工作量证明）、专为客户端指纹评估设计的多层检测，以及对超文本传输协议 (HTTP) 和传输层安全 (TLS) 特征的分析。

AI 僵尸网络的抓取流程

尽管基本的网络抓取类爬虫程序采用的抓取方法较为一致，但 AI 僵尸网络可以发现并抓取那些格式或位置不那么一致的非结构化数据和内容。不仅如此，AI 僵尸网络还能运用实际的商业情报来优化其决策过程。在表 3 的服务级别 3 中提及的复杂 AI 僵尸网络采用了三个步骤来抓取数据。这些步骤包括收集、提取，然后处理数据（图 5）。

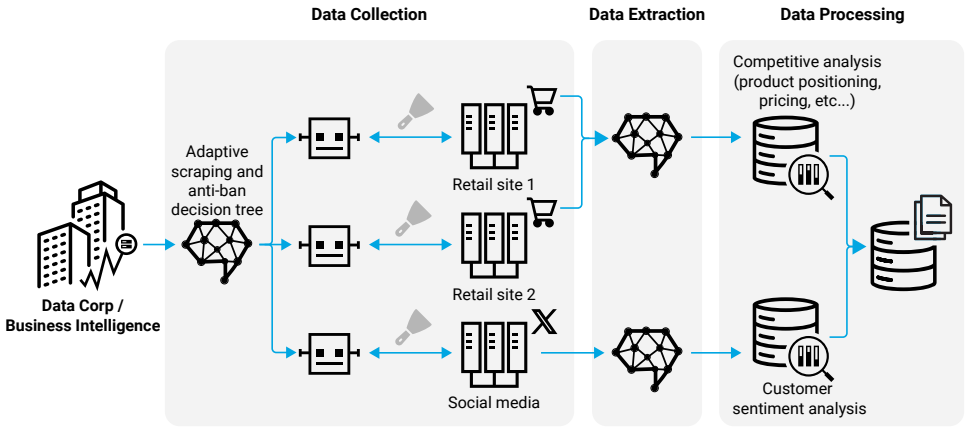

图 5：AI 僵尸网络及其三步流程的示意图

为了更全面地了解具体过程，我们详细探究一下这三个步骤。

数据收集

网络内容抓取涉及从一个或多个网站中提取数据，随后将这些数据整合以生成新的数据集。然后，企业可以根据需要应用和分析这些新数据集。所以，第一步是收集数据。

为了确保数据收集流程既迅速又顺畅，它需要将自适应抓取技术与“反禁令”或“反爬虫程序检测”技术相融合。这些技术构成了一个决策树，可全面检测任何潜在的保护机制。对于这一步，韧性是关键。爬虫程序防范措施可能包含 JavaScript 指纹识别、HTTP 和 TLS 指纹识别（用于评估 HTTP 标头和 TLS 握手），以及互联网协议 (IP) 声誉检测（图 6）。其中一些工作流程可能包含机器学习 (ML) 技术，特别是在收集成功率统计数据时；根据 cookie 策略、HTTP 标头和 TLS 参数进行调整时；以及评估 JavaScript 指纹识别代码时。这些环节也是无界面浏览器可以大显身手的地方。

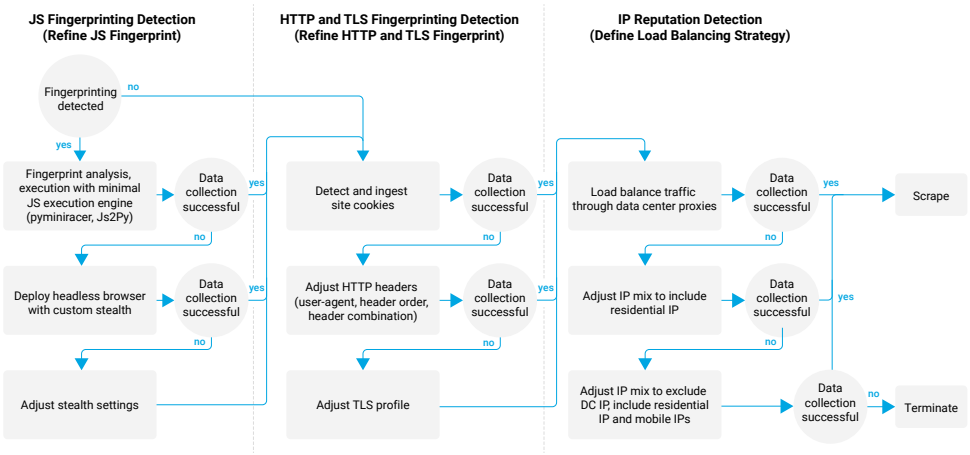

图 6：在尝试收集数据时，这个反爬虫程序检测决策树尝试规避 JavaScript 指纹识别、HTTP 及 TLS 指纹识别，以及 IP 声誉检测

没有界面的浏览器

无界面浏览器是一种没有图形用户界面 (GUI) 的 Web 浏览器。这意味着人类用户无法直接与无界面浏览器所呈现的网页进行交互，而是需要通过命令行界面 (CLI) 或者通过网络通信与浏览器进行交互。Selenium 是一款备受欢迎的开源无界面浏览器，常被用于自动抓取网页内容。对于尝试抓取动态内容的数据搜索者来说，这是一个非常有用的工具。

无界面浏览器不仅支持高效地截取网页截图和复制网站代码，更能在无需渲染整个页面的情况下精准提取所需数据。然而，实施无界面浏览器攻击的成本非常高，此类攻击有可能因其所留下的指纹而被检测到。但是，其他复杂基础架构的费用与这些无界面浏览器一样，都非常地高。

数据提取和数据处理

提取的信息主要包含 HTML 和 JSON 内容。在提取的所有数据中，仅有很少一部分对分析是有用的。例如，竞争分析通常涵盖价格、折扣信息、库存状态、产品 SKU 号、类别以及描述。ML 模型能够自动提取关键的信息片段。这些模型可以使用多种结构和数据格式进行训练，从而识别这些信息。这一工具能够省去手动提取数据时繁琐的额外处理工作，而且无需研究 HTML 和 JSON 内容代码结构。更重要的是，由于网站设计的不断发展，内容代码结构也会发生变化。当分析范围涵盖多个网站时，处理流程也必须使用额外的机器学习逻辑。

案例研究：网络内容抓取检测解决方案的优势

Akamai 研究人员对一部分受到网络内容抓取解决方案（能够检测抓取活动）保护的电商客户进行了观察，然后分析了这些客户一周内的流量活动细分数据。这相当于大约 69 亿次请求的样本大小。该分析仅考虑 HTML 和 AJAX 请求。由于大多数爬虫程序并不请求静态内容（如图像、JavaScript 和样式表），因此这些静态内容并未纳入分析范围；这一排除还有助于避免不必要的数据膨胀。

Akamai Content Protector 对整个活动进行了分类，其中 49.3% 是低风险人类流量，42.1% 是爬虫程序流量（27.5% 的高风险恶意爬虫程序和 14.6% 的良性爬虫程序），以及 8.7% 的中风险未分类流量（图 7）。

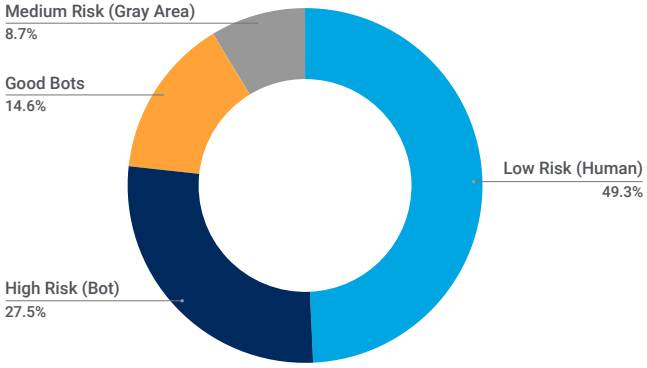

图 7：流量活动分类分解情况

图 8 展示了在总计 42.1% 的爬虫程序流量中，有高达 65.3% 的流量源自那些被认为是恶意的抓取类爬虫程序，而余下的 34.7% 则来自被归类为良性的抓取类爬虫程序（例如，网络搜索引擎、SEO、社交媒体以及在线广告服务）。

图 8：良性爬虫程序流量与恶意爬虫程序流量的比较

研究人员还衡量了高风险恶意爬虫程序（其在爬虫程序总流量中的占比高达 65.3%）的复杂程度。其中的 37% 源自基本的脚本化僵尸网络，能够通过简单的无状态方法轻松检测出来，47.6% 来自于更高级的脚本化僵尸网络，需要运用机器学习技术进行更高级的状态检测，15.5% 来自无界面浏览器，需要使用高级 JavaScript 指纹识别和状态检测方法（图 9）。

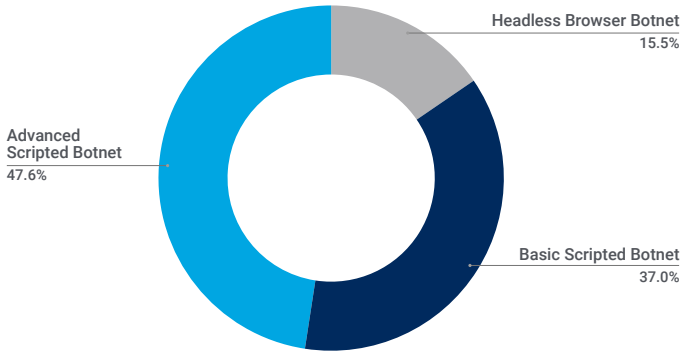

图 9：根据复杂程度划分的恶意爬虫程序流量分布（由于四舍五入，总数不完全等于 100%）。

从上述数据分析中，我们不难发现，恶意抓取类爬虫程序的数量远超过良性抓取类爬虫程序。此外，总流量中几乎有一半来源于爬虫程序，而在这其中，由高级脚本化僵尸网络产生的恶意爬虫程序流量占据了最大的比例 (47.6%)。

当针对这些爬虫程序的防御策略得以实施，成功移除这些抓取类爬虫程序后，网站的运行将更为迅速和高效，同时网站的各项指标也将变得更为清晰和准确。这些成果将为用户/客户带来更加优质的体验。如图 10 所示，采取抵御措施后，高风险爬虫程序请求的数量显著减少。

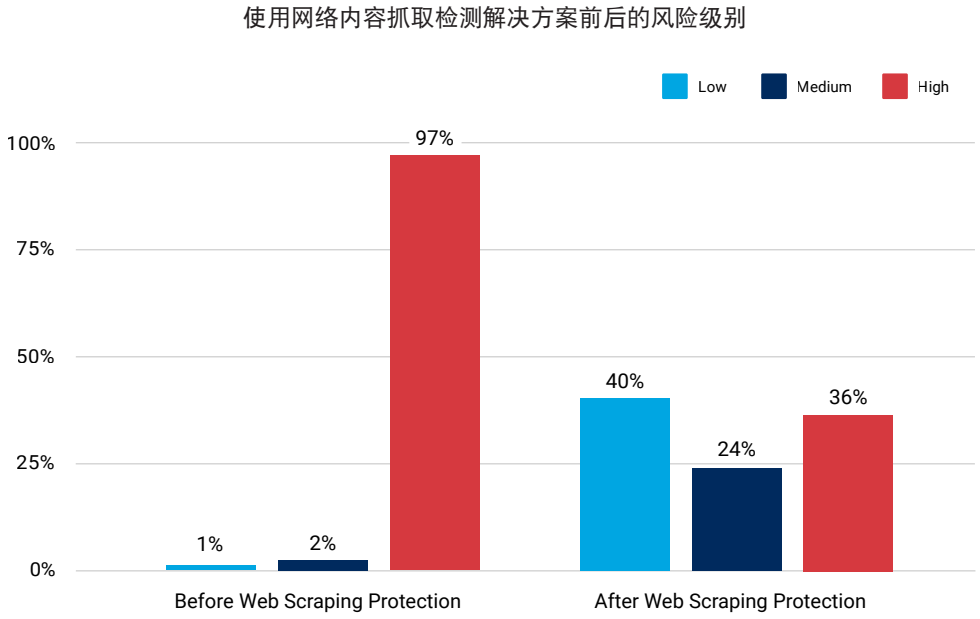

图 10：使用 Content Protector 进行抵御前后的风险级别

增强防护，抵御恶意爬虫程序

本部分介绍了检测网络抓取类爬虫程序的一些核心指标，并提供了针对这些爬虫程序实施防御措施所需工具的信息。

检测基本抓取类爬虫程序

尽管复杂的抓取类爬虫程序难以检测，但爬虫程序管理解决方案能够防御各类入侵式抓取类爬虫程序收集数据，在检测较为简单的网络抓取类爬虫程序时，尤其注重以下特征：

- 通告浏览器和操作系统版本较旧的请求
- HTTP 标头签名异常
- 使用旧版本 HTTP（例如 v1.1），而不是更常见的 HTTP v2 或者新兴的 HTTP v3
- 来自数以千计云服务/数据中心的请求

检测更高级的抓取类爬虫程序

对于较高级的抓取类爬虫程序，我们无法观察到上述列表中的任何特征。所以，这里有一些更复杂的抓取类爬虫程序的特征：

- 来自最新浏览器和操作系统版本的请求
- HTTP 标头设置看想来和合法浏览器的相同
- 使用 HTTP v2
- 来自数以千计住宅和移动 IP 地址的请求

识别流量模式

通过一些关键指标，我们可以识别网站产生的流量是来自人类访客（图 11）、基础爬虫程序（图 12），还是复杂的爬虫程序（图 13）。

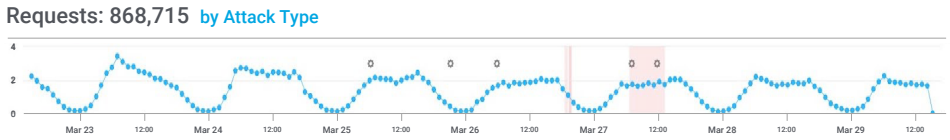

图 11：合法的用户流量通常显示出昼夜波动的活动周期

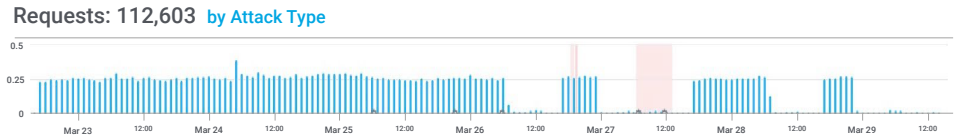

图 12：典型的爬虫程序流量通常表现为有规律的活动，伴有偶尔的间歇

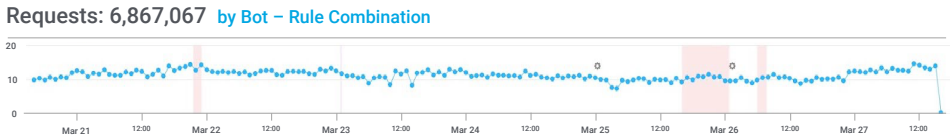

图 13：更复杂的爬虫程序会表现为流量全天候不间断

我们还常常观察到处于中间地带的僵尸网络，它们采用较弱的负载平衡策略，但却拥有复杂的指纹识别策略（反之亦然）。然而，更为先进的僵尸网络可能非常狡猾，它们拥有完美的指纹，甚至模拟合法的用户流量模式，从而通过检测。

除了需要警惕那些抓取类爬虫程序，利用 Content Protector 等工具来防范网络内容抓取，可能会为您带来额外优势，让您在抓取类爬虫程序频繁活动的网络环境中更加游刃有余。这些优势包括：

- 提高转化率和降低 IT 成本
- 提供更准确的指标，引导您做出更明智的投资决策，推动收入增长
- 降低价格压力，有助于您在竞争对手降价时保持销量稳定
- 保证客户能够买到心仪的商品，提高客户满意度。除此之外，客户在买到心仪的商品后，很可能会另外买一些其他的商品，从而带来额外的追加销售收入
- 保护品牌声誉，确保客户免受劣质假货的侵害，这些假货被误认为是原卖家的正品
- 保护产品收入，维护客户忠诚度
- 增加/保护广告收入
- 保留受众和网站访客

合规考虑因素

支付卡行业数据安全标准 (PCI DSS) v4.0 现已生效，许多改变都是由仍然对公司构成挑战的威胁趋势所驱动的。监测能力是应对这些攻击的关键所在。无论是在传统的 JavaScript 环境中，还是用于促进转型的 API 中，迅速检测和修复这些攻击都至关重要。

我们还注意到，在最新的 NIST 网络安全框架 2.0 版本中，出现了一个新兴的合规趋势，这一版本增加了治理功能。NIST 通常是许多政府法规的基础，并渗透到了许多商业网络安全框架之中。因此，现在是了解新指南的绝佳时机。您可以借助它来更新您的政策或将您当前的文档与之对照，了解是否有任何遗漏。

对于上市公司以及遵循公认会计原则 (GAAP) 的企业而言，网络安全重要性已成为另一个合规领域。为定义重大风险与威胁，需要领导团队展开协作。一旦识别出重大威胁（例如勒索软件），您需要应用相应的抵御措施（如微分段）。确保危机管理计划中包含了披露时间表，并预备好应对最坏情况的行动手册，在这种情况下，您需要向证券交易委员会提交网络事件表 8-K。

结论

我们希望这份报告能够对您有所启发，助您发掘那些可能对企业造成经济损失的领域。考虑将有越来越多的爬虫程序影响您的网站，优化良性爬虫程序，抵御恶意爬虫程序，并确保在整个客户体验过程中保持顺畅无阻变得至关重要。这是一个关乎业务发展的安全问题。在处理所有安全问题时，首先您需要获得监测能力；然后，分析安全问题对业务的影响；最后，确定风险与收入之间的 ROI，从而采取恰当的安全控制措施。

我们无法保护看不见的资产，因此现在是时候找出您的监测漏洞了。为实现这一点，您首先需要明确您网站上网络内容抓取活动的级别以及这些活动的意图。良性爬虫程序和恶意爬虫程序共同构成了爬虫程序环境，而抓取类爬虫程序根据其具体用途也有善恶之分。尽管良性和恶意抓取类爬虫程序之间的界限可能并不清晰，但爬虫程序的复杂程度却在不断演进（例如，执行无界面浏览器攻击的网络抓取类爬虫程序）。随着抓取类爬虫程序的活动，电子商务实体面临着 IT 成本上升和客户体验下降的重大挑战。因此，确保您拥有合适的工具来分析爬虫程序活动及其对您网站的影响变得至关重要。

您肯定不希望攻击者将他们的犯罪业务模式施加在您的网站上，并进行一系列的恶意活动，比如滥用忠诚度积分、发起欺诈性订单，甚至进行退货欺诈。同样，您也不希望看到票务爬虫程序抢夺限量供应的活动门票，或是购物爬虫程序抢购热门商品。爬虫程序可能会被恶意利用，通过肆意开设新帐户以获取特殊优惠，这种滥用行为会对营销活动的分析和成本造成负面影响。大型分布式拒绝服务 (DDoS) 僵尸网络对 Web 应用程序发动猛烈攻击，严重影响用户体验，让用户无法顺利下单或预订，进而导致收入损失和客户不满。有些爬虫程序甚至还能模拟人类在线行为，以增加网站点击量和流量，从而扭曲了对经过精心设计的数字化体验的营销和性能分析。您肯定不希望被这些爬虫程序所困扰。

正如我们先前所提及的，全球商业网络流量中超过一半是爬虫程序产生的，且爬虫程序的流量水平还在持续增加。Akamai 在这份报告中基于我们的安全平台提出了见解和建议，包含针对网络内容抓取的内容保护措施。我们与众多大型电子商务公司携手合作，希望通过分享各种保护措施和抵御策略，帮助广大公司更有效地保护其客户。我们预测，爬虫程序的使用量将会增加并会出现更多的服务级别，网络抓取类爬虫程序的类型也会更丰富。因此，您有必要定期评估公司的风险状况，确保当前的安全控制措施符合管理层的风险偏好。

敬请访问我们的安全研究中心，随时了解我们的最新研究资讯。

方法

Content Protector 数据

此数据样本说明了我们的 Content Protector 工具为其所监控的流量分配的风险级别分类。这些分类旨在检测爬虫程序的抓取活动，从而区分出良性爬虫程序与恶意爬虫程序。由于大多数爬虫程序并不请求静态内容，因此本分析只考虑 HTML 和 AJAX 请求，以避免不必要的数据膨胀。

这个数据样本涵盖了从 2024 年 4 月 12 日至 4 月 19 日这一周的时间段。我们的总样本大小包括超过 65 亿次请求。

爬虫程序攻击

这些数据表示通过我们的 Web 应用程序防火墙 (WAF) 和爬虫程序管理工具观察到的流量的应用层告警数量。如果在针对受保护的网站、应用程序或 API 的访问请求中检测到爬虫程序负载，系统就会触发爬虫程序告警。恶意和良性爬虫程序都可能触发此类爬虫程序警报。警报并不表示攻击已经得手。虽然这些产品允许的定制程度极高，但我们在收集此处提供的数据时，所采用的方式并未考虑受保护资产的定制配置。这些数据来自一个内部工具，专用于分析在 Akamai Connected Cloud 上检测到的安全事件。Akamai Connected Cloud 是一个庞大的网络，在全球 130 多个国家/地区将近 1,300 个网络中的 4,000 多个地点拥有约 340,000 台服务器。我们的安全团队使用这些数据（每月达到 PB 级）来研究攻击，标记恶意行为并将其他情报馈送到 Akamai 解决方案中。

该数据涵盖了从 2023 年 1 月 1 日到 2024 年 3 月 31 日的 15 个月的时间段。

致谢名单

总编辑

Lance Rhodes

编辑与创作

David Senecal
Maria Vlasak

审稿和主题撰稿

Mitch Mayne
Susan McReynolds
Christine Ross
Badette Tribbey
Steve Winterfeld

数据分析

Chelsea Tuttle

推广材料

Annie Brunholzl

营销与发布

Georgina Morales
Emily Spinks

进一步阅读《互联网现状/安全性》报告

《互联网现状/安全性》报告由 Akamai 精心呈献，获得了各界的广泛赞誉。请前往以下网址回顾往期报告，并关注即将发布的新报告：
akamai.com/soti

进一步查看 Akamai 威胁研究

请前往以下网址，了解最新的威胁情报分析、安全报告和网络安全研究的动态：
akamai.com/security-research

访问此报告中的数据

查看本报告中引用的图片和图表的高画质版本。这些图片可供免费使用和引用，但必须注明转载来源，并保留 Akamai 徽标。
akamai.com/sotidata

进一步探索 Akamai 解决方案

如需了解有关 Akamai 检测和防范网络抓取类爬虫程序的解决方案，请访问 **Content Protector 页面**。