arXiv:2304.14359v1 [cs.CY] 26 Apr 2023

免费内容网络的测量与建模*

Abdulrahman Alabduljabbar[1],Runyu Ma[2],Ahmed Abusnaina[3],
Rhongho Jang[4],SongqingChen[2], DaeHun Nyang[5], 和 David Mohaisen[1]

[1] *University of Central Florida*、 [2] *George Mason University*、 [3] *Meta Inc*、 [4] *Wayne State University*、 [5] *Ewha Womans University*

摘要

多年来，互联网上一直存在提供免费书籍、音乐、游戏、电影等内容的免费内容网站。人们普遍认为，这类网站可能不同于提供相同内容类型的高级网站，但文献中缺乏支持这一观点的分析。尤其是，目前尚不清楚这些网站是否与其付费对应网站一样安全。本文旨在通过分析和量化，探究免费内容网站与高级网站之间的相似之处和差异，包括它们的风险概况。为开展这项分析，我们整理了一个包含 834 个提供书籍、游戏、电影、音乐和软件的免费内容网站，以及 728 个提供相同类型内容的高级网站的列表。随后，我们从域名、内容和风险层面展开分析，考察并比较这些网站的域名、创建时间、SSL 证书、HTTP 请求、页面大小、平均加载时间和内容类型。在风险分析中，我们考察并分析了这些网站在网站层面和组件层面的恶意性。除其他有趣发现外，我们发现，免费内容网站往往广泛分布于各个顶级域名中，且动态变化更多，新注册域名呈上升趋势。此外，免费内容网站使用过期证书的可能性是高级网站的 4.5 倍；在网站层面被判定为恶意的可能性是其 19 倍，在组件层面被判定为恶意的可能性是其 2.64 倍。

*本研究得到韩国国家研究基金会（NRF）通过科学技术信息通信部资助的全球研究实验室（GRL）项目（NRF-2016K1A1A2912757）的支持。A. Alabduljabbar还得到沙特阿拉伯文化使团（SACM）的部分支持。R. Ma和S. Chen还得到Commonwealth Cyber Initiative资助项目以及NSF资助项目CNS-2007153的部分支持。（*corresponding authors: David Mohaisen; mohaisen@ucf.edu.*）。 η共同贡献者

鉴于这两类网站之间存在明显差异，我们进一步探讨了免费内容风险网站风险建模的自动化和泛化能力，并表明一种简单的基于机器学习的技术能够以 86.81% 的准确率识别这类网站。

关键词： 免费内容网站；网络安全；网络挖掘

1. 引言

网站根据其盈利方式大致分为两类：免费内容网站和付费网站。免费内容网站免费提供内容，通常依靠广告收入和用户捐赠维持运营 [1, 2, 3, 4], ；付费网站则通过收费提供服务，例如订阅或按使用付费模式 [5]。付费网站凭借精心设计的网站，以及通过专门工程和运营工作进行的维护，确保很高的服务质量 [6]。相比之下，人们普遍认为，免费内容网站对服务质量的期望不高，而且通常由用户驱动 [7]。

对功能和安全质量要求宽松、由用户驱动的内容，以及免费内容平台上第三方广告的大量使用，都会带来各种风险 [8, 9, 10, 11]。例如，这些网站上的广告可能被利用来造成数据和信息泄露，甚至在用户设备上分发和执行恶意脚本 [12, 13]。此外，免费内容网站缺乏严格的维护运营规则，也会引发各种风险：这类网站使用的 Web 框架很少更新，使旧有的未修补漏洞有机可乘，并让其用户面临不同程度的风险 [14]。

然而，撇开未经检验的假设和广泛未经证实的信念不谈，提供相同类型内容的免费内容网站与高级网站是否存在差异？免费内容网站在结构、内容和安全性方面是否不同于高级网站

的属性？这些网站是否会给用户带来隐藏成本，从而超过其表面上的优势，即免费？为回答这些问题，我们对精心整理的数据集进行了系统分析，该数据集收录了834个免费内容网站和728个高级网站。我们的研究结合了域名级分析和内容级分析，并从多个维度开展安全分析。在域名级分析中，我们考察域名系统特征、创建时间和SSL（安全套接层）特征，以衡量网站意图。在内容级分析中，我们考察HTTP（超文本传输协议）请求、页面大小、加载时间和内容类型，这些指标均用于衡量网站复杂度。在安全分析中，我们使用两种主流的现成工具VirusTotal API和 $VirusTotal\ API$ [15] Sucuri API $Sucuri\ API$ [16]，同时考察网站级和组件级检测与漏洞。

我们的分析得出结论：提供相同类型内容的免费内容网站与高级网站之间存在显著、根本且内在的差异。在其他有趣的发现中，我们报告称，免费内容网站几乎完全广泛分布于顶级域名（TLDs）中，尽管使用的是常见的二级域名（SLDs）。此外，它们经常更换域名，更有可能规避列入黑名单，也更常与无效 SSL 证书相关联。在内容方面，免费内容网站通常只需更少的 HTTP 请求即可获得更小的请求页面大小，但代价是加载时间显著增加，因为它们大量采用重定向并包含更多脚本对象。在风险方面，我们发现，与高级网站相比，免费内容网站在页面级别（38% 对 2%）和文件级别（45% 对 17%）成为恶意网站的可能性分别高出 19 倍和 2.64 倍。

我们利用从这些分析中获得的洞见，通过对免费内容网站的风险进行建模来概括并推断其规律。为此，我们使用纯性能指标定义风险。此外，我们还能够以极高的准确率（超过 86%）对高风险网站进行分组。

贡献与发现。本文对不同维度上的同类内容免费网站和付费网站进行了深入的比较分析，涵盖

域名、内容和安全。借助功能丰富的分析方法，我们构建了一种基于机器学习的方法，能够以较高的准确率评估免费内容网站的风险。下文将详细介绍我们的贡献。

1. **免费内容网站整理（§3）**。我们整理了一份包含超过 1,500 个免费内容网站和提供相同类型内容的高级网站的列表。这些网站来自 Google、DuckDuckGo 和必应搜索引擎的顶级搜索结果。随后，我们对这些网站进行爬取，以获取其内容，包括脚本、图像、HTML、CSS 等。

2. **域名级分析（§4.1）**。为了考察免费内容网站的域名级特征，我们分析了三个方面：其顶级域名（TLD）、SSL证书和创建日期。结果显示，免费内容网站的数量显著增加，而新创建的高级网站数量则有所减少。此外，我们观察到，免费内容网站中的域名动态变化比高级网站更为频繁。近三分之一的免费内容网站使用了无效或未匹配的 SSL 证书。

3. **内容级分析（§4.2）**。为了考察内容级特征，我们分析了三个方面：HTTP 请求、页面大小和平均加载时间。除其他发现外，我们观察到，高级网站包含的图像明显更多，其平均大小是免费内容网站的三倍。然而，有趣的是，我们发现由于各种内在的设计选择，包括利用脚本和通过重定向投放广告（这些做法在免费内容网站中更为普遍），高级网站的加载时间似乎与免费内容网站相当。

4. **免费内容网站风险分析（§5.1）**。我们利用两种广泛使用的现成工具——$VirusTotal$ 和$Sucuri$——来评估与免费内容网站相关的安全风险。我们的分析显示，免费内容网站与恶意性的关联显著高于付费网站

。不过，发现被检测为恶意的付费网站相当有趣，值得进一步探究。

5. **风险建模（§5.2）**。性能和安全指标分析均凸显了免费内容网站与高级网站之间的显著差异。此外，两者的风险概况也截然不同。受其特征差异的启发，我们构建了一种简单的机器学习算法，利用易于获取的域名和内容级特征来预测网站的风险。对于免费内容网站的风险建模，我们取得了86.81%的准确率，结果颇具潜力。

2. 相关工作

与本文贡献最相关的工作主要分为两个方向：网站分析和恶意网页内容分析。下面将概述这两个方向的一些研究成果。

2.1. 网站分析

网站的内容和使用方式在不断演变，其组件的复杂性和丰富性也随之增加。然而，随着这种演变，由于各组件之间的相互作用，会产生各种安全风险 [17, 18, 19, 20]。文献中一个尚未得到充分研究的安全方面是网站证书的有效性 [21]。为解决这一问题，Chung 等 [21] 对 Web 公钥基础设施中的证书进行了深入分析，结果表明，Web 公钥基础设施中的绝大多数证书均为无效证书。他们还调查了无效证书的来源，得出结论认为，这些证书主要由终端用户设备生成，并且会定期重新生成新的自签名证书。

Libert等人 [22]评估了百万热门网站采用的侵犯隐私的做法，例如数据泄露。他们得出结论，约有十分之九的网站未经用户同意，便与第三方服务共享用户数据。Lavrenovs等人 [23] 使用类似的数据集，开展了一项全面的

对Alexa排名前百万网站的安全评估，结果显示，29.1%的HTTPS请求存在错误的TLS（传输层安全）配置，而仅有17.5%的网站实施了HTTP严格传输安全（HSTS）策略。这些发现令人担忧，表明此类热门网站所遵循的安全策略状况堪忧。

为了探索用于评估Web应用程序安全漏洞的环境，Alsmadi 等人 [24] 设计了一种针对各种无效输入的基于组件的测试机制，并利用该机制研究网站因这些输入而表现出的行为，包括安全性。由于无效输入是攻击面中持续存在的一部分，因此通过消除这些输入（即拒绝无效输入），可以增强在线服务和Web应用程序的安全性。为此，他们提出了多种检测无效输入的方法，并发现了许多SQL注入漏洞。

2.2. 恶意*Web*内容

近期研究表明，攻击者能够将恶意代码嵌入网站的*JavaScript*、*GIF*或*Redictection* 组件中 [25, 26, 27, 28, 29]。终端用户的安全性在很大程度上取决于对这类恶意内容的检测和防范，相关内容也已得到研究。为此，研究人员利用了Web应用程序的各种特征，包括URL（统一资源定位符）域组件、网页内容、HTTP头和已加载脚本，并使用这些特征检测恶意Web应用程序 [30, 31]。研究还表明，一个很有前景的特征集是HTTP头信息 [22], ，其中McGahagan 等人 [32] 利用其中672项特征构建了一个恶意网站检测系统。为检验使用组件和内容（即文件和脚本）作为检测特征的可行性，作者对不同的网页内容特征进行了全面评估。这些新设计的17项特征能够提升恶意网站检测性能。

网站一个至关重要但尚未被探索的方面，是其上投放的广告与相关恶意性之间的相互作用。Li 等人 [33]

研究了各种恶意在线广告和营销方法，例如恶意软件传播、
[34, 35, 36, 37, 38], 点击欺诈等。他们的研究使用了一个大规模广告相关网络轨迹数据集，展示了数百个高排名网站中的恶意广告行为。为检验恶意广告检测的有效性，Masri 等人 [39] 评估了三种工具：$VirusTotal$、$URLVoid$和$TrendMicro$，结果表明 URLVoid 的性能最佳。

另一项受到广泛研究的突出威胁是免费软件下载门户上恶意内容的传播 [40]。此类门户可能被恶意利用，将有害软件分发至终端用户设备。Rivera等[41]对通过免费软件下载门户获取的潜在有害程序（PUP）和恶意软件进行了系统分析，结果表明，下载内容中平均有8%至26%属于潜在有害程序或恶意内容。

机器学习算法也被广泛用于有效检测恶意网站 [42]。然而，它们受到两个关键挑战的限制：特征选择和规避。为解决特征选择问题，Singh和Goyal在其分析中主张将特征选择与额外性能和准确率相结合 [43]。检测规避是另一个问题，通常与固有特征相关，包括使用重定向和隐藏式内嵌框架。在该领域，Liu和Lee [44] 提出了一种基于卷积神经网络的有效恶意内容检测方法，该方法基于网页的屏幕截图。

本研究。在本研究中，我们探讨并评估免费内容网站的恶意性，并将其与付费网站进行对比。研究结果显示，免费内容网站中恶意内容所占比例呈令人担忧的上升趋势。为主动应对这一问题，我们利用易于获取的性能特征对这些网站的风险进行建模，并识别出高达 86.81% 经真实标签验证的高风险网站。

表1：所收集数据集的概览。收集的 URL 涉及五个不同类别，来自免费内容网站和付费网站。总体而言，本研究共爬取了 1,562 个网站。

类别	免费内容网站			付费网站		
	URLs	文件	平均文件数	URLs	文件	平均文件数
书籍	154	7,073	45.93	195	17,840	91.49
游戏	80	6,439	80.49	113	11,314	100.12
电影	331	9,821	29.67	152	10,738	70.64
音乐	83	6,059	73.00	86	7,225	84.01
软件	186	11,561	62.16	182	18,742	102.98
总体	834	40,953	49.10	728	65,859	90.47

3. 数据集概述

接下来，我们将重点介绍创建数据集所采用的方法，包括初步筛选及相关标准、人工标注、爬取和数据增强。

网站选择。 我们整理了一份包含 1,562 个免费内容网站和付费网站的清单，用于开展测量。选择网站时，我们考虑了多种代表性限制。具体而言，选择网站时采用了以下标准：

(1) 受欢迎程度： 每个网站都必须是网络上最受欢迎的网站之一。鉴于这些网站未必都是最受欢迎的网站，我们使用搜索引擎的结果作为估算其受欢迎程度的依据。如果某个网站出现在至少一个所用搜索引擎（Google、DuckDuckGo 和必应）的靠前搜索结果中，就会被视为受欢迎。

(2) 均衡表示： 在编制总体数据集时，我们确保数据集在每个类别中都保持均衡。为此，我们不断扩展查询，直到免费内容网站和付费网站在各类别中都达到接近均衡的表示。

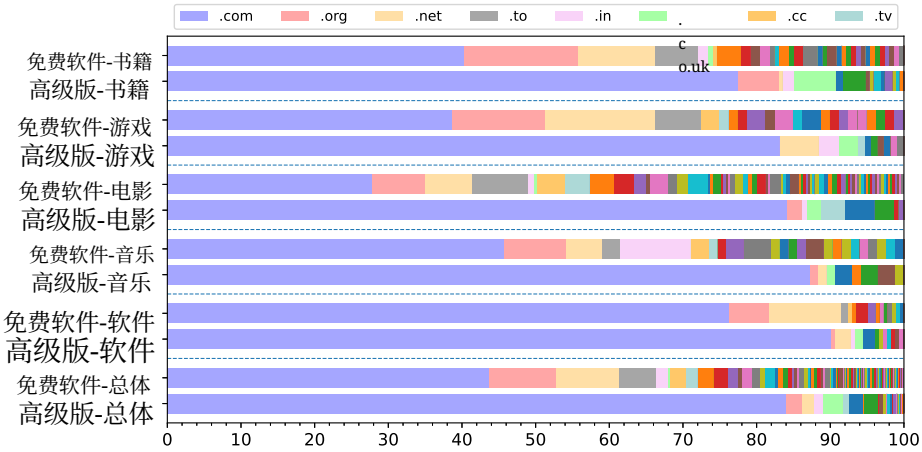


图1：免费与付费网站的顶级域名分布。与付费网站相比，免费内容网站分布在更多不同的顶级域名下。

最初选出要纳入数据集的独特网站后，我们会继续手动检查并标注每个网站，将其归类为高级版或免费内容网站。随后，我们也会根据网站主要提供的内容类型，将每个网站手动归入五个类别之一：书籍、游戏、电影、音乐或软件。

网站爬取。 为了解免费内容网站相关的风险，我们使用 PyWebCopy [45]，（一种用于克隆网站并下载其相关文件的 Python 软件包）爬取了每个网站的内容（即文件）。随后，我们使用获取的文件进行风险分析和建模，因为这些文件反映了所提供服务的行为。然后，我们为数据集补充了多种属性，并将其分为两大类：域级属性（顶级域名、域名创建信息、SSL 证书信息）和内容级属性（HTTP 请求信息、页面大小、加载时间和内容类型）。

高层次特征。 表1展示了所收集数据集的分布情况。请注意，从高级网站抓取的平均文件数显著高于从免费内容网站抓取的平均文件数。

4. 网站分析

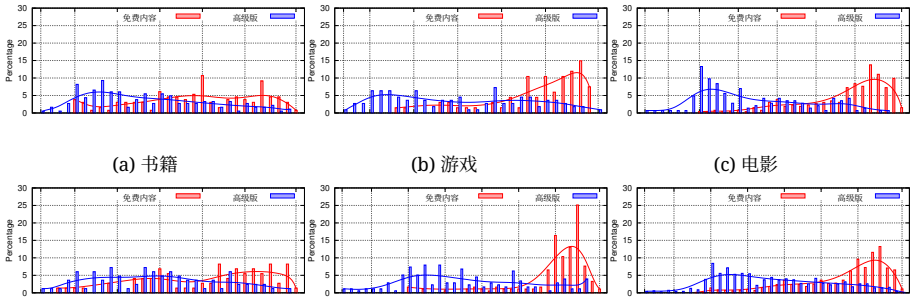
为了了解免费内容网站与高级网站之间的根本差异，我们进行了两类分析：域名级分析和内容级分析。域名是网站的入口，其中包含丰富的信息，可用于了解网站的意图。将域名级特征与内容级特征结合起来，可以更深入地了解网站的意图。下面我们将基于这两组特征给出分析结果。

4.1. 域名级分析

域名级分析从所有者信息、创建日期和所使用的顶级域名等方面，为我们提供了对网站作为基础设施的宏观且有趣的视角和理解。我们开展此类分析，是为了对比免费内容网站和高级网站所关联的域名。

顶级域名分析。 顶级域名是分层域名系统中级别最高的域名，其下是二级域（Second-level Domain）；在 example.com 中，example 是二级域，而 com 是顶级域。近年来，随着新的通用顶级域（gTLDs）的引入，顶级域名的数量显著增加，尽管 *'.com'*、*'.net'*、*'.org'* 和 *'.edu'* 仍然最为突出 [46]。在本研究中，我们考察了免费内容和高级网站在各顶级域名中的分布情况，如图1所示。

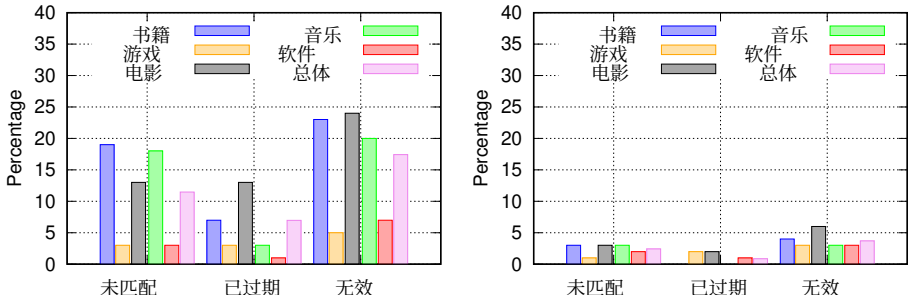
我们发现，*'.com'* 是最主要的顶级域名，分别有44%的免费内容网站和84%的付费网站使用了*'.com'*。不过，有趣的是，我们发现数据集中付费网站使用的唯一顶级域总数仅为24个，而免费内容网站中的这一数字为98个域名。我们注意到，这种广泛分布可能是由主流浏览器和系统所采用的恶意网站拦截机制触发的。例如，Chrome和Firefox在使用安全浏览服务 [47]时依靠用户报告来收集并拦截恶意网站。为规避拦截，免费内容网站会定期更换其域名



(a) 书籍 (b) 游戏 (c) 电影

(d) 音乐 (e) 软件 (f) 总体

图2：免费内容网站与高级网站的域名创建年份对比。通过比较不同内容类型的趋势，我们发现，与高级网站相比，免费内容域名创建呈现显著的上升趋势。

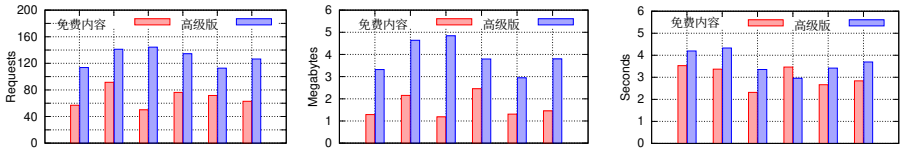


（a）免费内容网站 （b）高级网站

图 3：SSL 证书分析结果。我们发现，近36%的免费内容网站存在 SSL 证书问题，而高级网站中这一比例为7%。

名称。然而，免费内容网站运营者会保留相同的二级域，并将网站迁移到其他顶级域名，以留住现有用户；其中一些运营者还会更换顶级域名以规避拦截。

域名创建。 我们考察了网站创建日期，发现新创建的免费内容网站数量呈上升趋势，而新创建的付费网站数量则呈下降趋势，如图2所示。这一增长趋势，尤其是在2015年至2021年期间，促使我们进一步考察并了解使用在线免费内容网站所涉及的风险。为进一步支持这一点，我们通过对顶级域名的分析发现



(a) 每页的 HTTP 请求数。 (b) 大小（MB）。 (c) 加载时间（秒）。

图 4：免费内容网站与付费网站的页面相关对比（平均统计数据）。尽管页面大小不同，免费内容网站与付费网站的平均页面加载时间相近，这表明影响加载时间的因素并非只有页面大小。

免费内容网站往往会定期更换其域名，以避免内容屏蔽或被列入黑名单。

SSL 证书分析。HTTP 将网站内容（如 HTML）从 Web 服务器传输到用户浏览器。然而，这种协议并不安全，传输的数据可能会遭到未授权访问。因此，大多数网站已转而使用安全版本的 HTTP（HTTPS），该协议采用加密机制来保护传输的内容。我们的分析发现，36% 的免费内容网站使用的 HTTPS 无效，而付费网站中这一比例仅为 7%。此外，我们发现，26% 的免费内容网站仍允许通过 HTTP（不安全）访问，而付费网站中允许 HTTP 访问的比例为 0%。SSL 证书是一种数字身份验证方法，用于验证网站身份，并为 HTTPS 提供服务器与客户端计算机之间的加密连接。SSL 证书是网站保护用户数据及防范网络钓鱼等威胁的关键组成部分。

在本研究中，我们调查了免费内容网站和付费网站的 SSL 证书有效性。具体而言，我们研究了三个方面：(i) 证书中的主机名未匹配，(ii) 证书已过期，以及 (iii) 证书无效或伪造。图 3 显示，免费内容网站中总计有 36% 的网站证书存在问题（即 11.5% 的名称未匹配、7% 的证书已过期，以及 17.5% 的证书无效），而

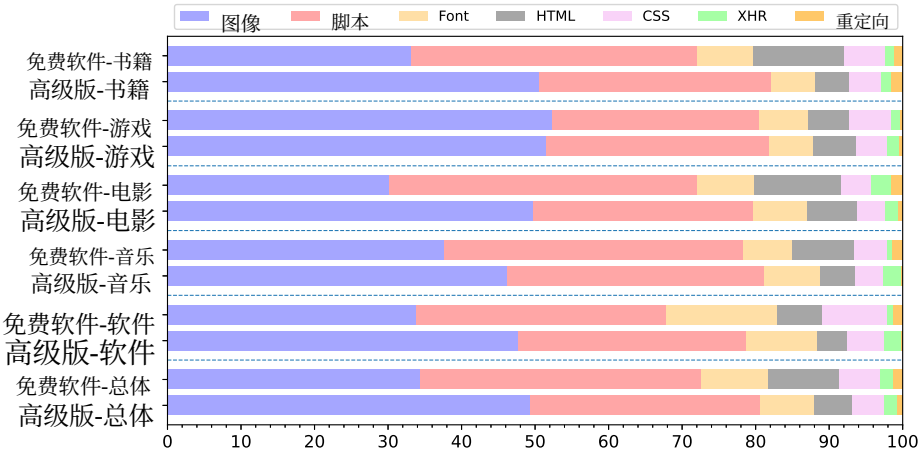


图 5：免费内容与付费网站之间的内容类型比较。我们观察到网站文件类型存在一些差异，尤其是在 *images*类型方面。

付费网站中关联的 SSL 证书存在问题的比例总共仅为 7%。在“电影”、“书籍”和“音乐”类别中，这种情况更加明显。如图所示，免费内容网站的 SSL 证书更容易出现问题。这可能是因为免费内容运营者不愿增加运营成本，因而没有续签 SSL 证书。尽管如此，这种做法仍会给用户信息和数据隐私带来潜在风险。

要点：通过域名级分析，我们发现：(i) 免费内容网站更为新近，近年来其数量显著增长；相比之下，高级网站的增长正在放缓，每年新出现的网站越来越少；(ii) 免费内容网站的域名分布在更多顶级域名上，因为它们会更换域名以规避恶意网站屏蔽机制；(iii) 免费内容网站更有可能使用无效或已过期的 SSL 证书。这些发现补充了我们对使用免费内容网站的安全性及其相关风险的分析。

4.2. 内容级分析

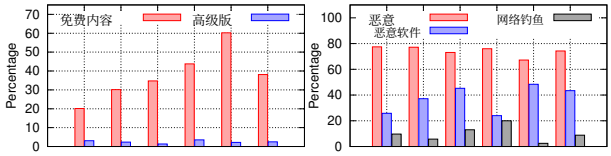
为深入了解免费内容网站和高级网站的内容级特征，我们分析了这两类网站中提取的文件。在此分析中，我们重点关注四项特征：HTTP 请求数量、页面大小、页面加载时间和内容类型。

HTTP 请求。 客户端向服务器发出 HTTP 请求，以请求访问资源（例如 HTML 文件、CSS、图像），每个页面的请求数量可以反映所请求页面的复杂程度。图 4a 展示了免费内容网站和付费网站发出的 HTTP 请求平均数量。我们观察到，与访问免费内容网站相比，客户端访问高级网站时发起的请求数量几乎是前者的两倍。鉴于高级网站页面的页面大小更大，这一结果并不意外。然而，我们观察到，高级网站的平均页面大小是免费内容网站的 3 倍，而 HTTP 请求数量仅多出 2 倍，这表明，访问免费内容页面需要为相同数量的数据发出更多 HTTP 请求。这可能是重定向造成的，因为每次重定向都会触发一个或多个独立的 HTTP 请求，并且会消耗更多加载时间，后文将对此进行说明。

平均页面大小。根据 HTTP Archive [48],发布的页面重量报告，排名前一百万的网站平均页面大小约为 2.07 MB。属于这两个类别的平均页面大小相差多少？为回答这个问题，我们考察了免费内容网站和高级网站的平均页面大小，结果见图 4b。我们发现，免费内容网站的页面大小符合 HTTP Archive [48], 报告的正态分布，而高级网站的平均主页大小为 3.9 MB，是免费内容页面平均大小的三倍。一种可能的解释是，正如我们稍后将展示的，免费内容网站依赖于将用户重定向到其他网站的内容或广告网站，而不是在免费内容页面正文中包含并呈现内容。

平均页面加载时间。我们将页面加载时间定义为页面完全加载所需的时间，并对其进行测量，以了解网站复杂性的其他方面。图 4c 展示了使用 SolarWinds Pingdom API（应用程序编程接口） [49],计算得出的免费内容网站和高级网站的平均页面加载时间。尽管高级网站的平均大小是免费内容页面平均大小的三倍，但我们发现两者的平均加载时间相近，这表明除页面大小外，还有其他因素会影响加载时间，例如性能下降和频繁使用重定向。

内容类型。 页面大小似乎无法完全解释网站的复杂程度和加载时间，因此有必要深入分析网站内容。内容类型是网站内容在组件层面（即文件）的另一项统计特征。这些组件包括图像（*GIF*、*PNG*、*JPEG*）、*JavaScript*、文本、*HTML*、*CSS*、*XHR*和重定向。我们发现，图像 是最常见的组件，其次是 *JavaScript*，而重定向 内容在这些组件中最少。图 5 展示了免费内容网站和高级网站中不同组件的平均分布比例（%）。总体而言，高级网站的图像比免费内容网站多 15%。然而，我们注意到免费内容网站大量使用重定向 ，因为这通常是投放广告并误导过滤算法的一种方式。我们发现，免费内容网站与高级页面的重定向比例（四舍五入后）分别为 6（软件）、7（音乐）、3（电影）、1（游戏）和 1（书籍）。总体而言，免费内容网站的重定向次数是高级网站的两倍，其 HTML 内容是两倍，CSS 内容是 1.5 倍，JavaScript 内容是 1.23 倍。



（a）标记为恶意。　　（b）恶意免费内容网站。

图 6：免费内容网站和高 级版网站的潜在恶意性。

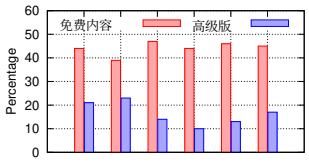


图 7：　　恶意文件由 *VirusTotal*检测到：免费内容与高级版。

要点： 我们的内容级分析揭示了免费内容网站与高级网站之间的主要差异。我们发现：（i）高级网站的请求数几乎是免费内容网站的两倍，页面平均大小是免费内容网站的三倍。（ii）尽管如此，免费内容网站与高级网站的主页加载时间相近。（iii）从内容类型来看，免费内容网站中重定向 组件所占比例更高，因为这是投放广告的主要方式。

5. 恶意性分析

我们目前开展的分析考察了免费内容网站和付费网站的性能及非安全特征，凸显出二者之间的明显差异，而这些差异会导致直接和间接成本。衡量免费内容网站成本的一个最重要且显而易见的指标，是了解其安全性及相关风险。在本节中，我们将开展这项分析，重点关注威胁指标，例如 URL、文件的恶意性及相关漏洞（§5.1）。为实现此类风险的自动化发现，我们还报告了一种基于机器学习的工具的结果。该工具根据从风险分析中获得的特征，展示网站的风险边界（§5.2）。

5.1. 风险评估

研究服务网站的恶意性和漏洞，考察它们可能如何影响用户体验、安全性和安全，这一点十分重要。基于此，我们使用以下几项指标定义网站风险，即：（i）包含恶意软件；（ii）运行恶意脚本；（iii）利用用户设备资源；或（iv）包含漏洞、过时的软件版本或未修补的框架。

为评估每类网站的风险，同时避免重复造轮子，我们利用两个公共 API：$VirusTotal$ [15] 和$Sucuri$ [16]，对有害行为进行分析。$VirusTotal$ 是一项在线服务，可汇总 70 多个扫描引擎的扫描结果，既可用于扫描文件，也可用于扫描 URL。另一方面，Sucuri 是一项服务，可检测网站是否存在多种已知恶意软件、病毒、黑名单列表、漏洞、过时框架和恶意代码。

恶意 URL 检测与标注。 通过使用$VirusTotal$ API，我们提取了与网站 URL 相关的恶意活动，如图 6 所示。我们注意到，免费内容网站与付费网站在恶意性方面存在明显差异。具体而言，图 6a 显示，38% 的免费内容网站被$VirusTotal$认定为恶意网站，而付费网站中这一比例仅为 2%。在这些已检测到的网站中，有相当一部分（≈74%）被标记为恶意网站（图 6b）；这类网站旨在推广诈骗、攻击和欺诈。我们还注意到，相当一部分免费内容 URL 被检测为恶意 URL，比例从 20%（“书籍”网站）到 60%（“软件” 网站）不等。相比之下，付费网站的检测率非常低，仅为 1% 至 4%。

恶意文件检测与格式分析。 为了了解给定服务（即内容提供商）的行为，有必要分析该服务托管的可执行脚本的行为特征。这些脚本会以文件形式转发给终端用户，包括图像、JavaScript 代码、HTML 以及其他格式，并且通常会在

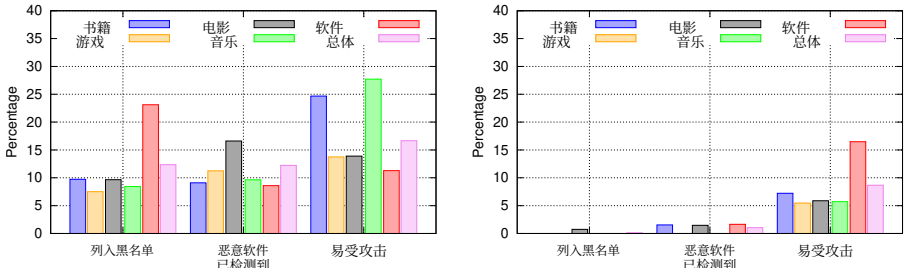
表 2：免费内容和付费网站中不同文件格式的恶意文件分布。我们观察到，在这两种情况下，都有很大一部分 *".gif"* 文件被标记为恶意文件，尽管免费内容中的比例几乎是其两倍。

	类别	.gif	.html	.png	.js	.php	.woff	.jpg	.eot	.woff2	.svg	.ttf	.log	.css
Free Content	书籍	28%	1%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	游戏	7%	13%	0%	1%	3%	0%	0%	0%	0%	0%	0%	0%	0%
	电影	40%	6%	1%	2%	0%	1%	0%	1%	0%	1%	1%	1%	0%
	音乐	26%	6%	0%	1%	4%	3%	0%	3%	3%	0%	2%	0%	0%
	软件	11%	30%	4%	1%	1%	4%	0%	2%	4%	2%	3%	1%	0%
	总体	26%	11%	2%	1%	2%	2%	0%	2%	4%	1%	2%	1%	0%
Premium	书籍	19%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	游戏	21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	电影	9%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	音乐	21%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	软件	5%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	总体	15%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

用户设备上呈现或执行。分析脚本和网站文件至关重要，因为近期研究 [50, 51, 52, 53] 表明，此类内容可能会被利用，从而导致信息和数据泄露，并滥用终端用户设备的资源。为了了解免费内容网站的风险，我们利用 *VirusTotal* API 进行恶意文件识别。相比之下，图 7 展示了 *VirusTotal* API 在免费内容网站和高级网站中检测到的恶意文件百分比。虽然属于高级网站且被标记为恶意的 URL 数量仅为 2%，但其被标记为恶意的文件数量达到了 17%。

我们发现，整体趋势依然存在，而且差异更为显著：45% 的免费内容网站含有被标记为恶意的文件（高级版网站中这一比例为 17%）。为更好地理解这一现象，我们调查了恶意文件的格式分布，比较结果见表 2。

根据表 2，我们报告称，大多数恶意文件采用*'.gif*和*'.html'* 格式。这是因为：(i) *'.gif'* 和*'.html'* 文件



(a) 免费内容网站　　(b) 付费网站

图 8：评估免费内容和付费网站的恶意性。我们展示被标记为列入黑名单、恶意软件和易受攻击的网站所占的百分比。

包含恶意嵌入式脚本，或者 (ii) VirusTotal 引擎通常将*'.gif'* 文件视为含有恶意内容（即可能是假阳性）。值得注意的是，我们手动检查了*'.gif'* 文件，并发现大多数被标记为恶意的*'.gif'* 文件都是与广告相关的内容。

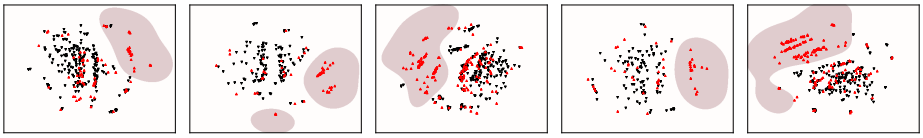
网站漏洞与黑名单。为了分析潜在的可利用漏洞和列入黑名单的情况，我们利用 Sucuri API [16] 获取这两类服务的域名活动信息。结果发现，如图 8 所示，12% 的免费内容网站被检测到含有恶意软件，而高级版网站中这一比例仅为 1%。此外，我们发现免费"电影"网站的恶意软件检测率最高（16.67%），如图8a所示。

我们还扫描了这些网站以查找漏洞，发现免费"书籍"和"音乐"网站的总体漏洞数量最多。尽管付费网站的漏洞报告率较低，仍有 17% 的"软件"被标记为易受攻击，这一比例高于免费内容网站（12%），相当令人意外。根据*Sucuri* 的报告，合法"软件"网站中很大一部分漏洞是由过时的框架版本导致的，这在"软件"服务网站中很常见。

就列入黑名单而言，图8a显示，12%的免费内容网站被*Sucuri* 扫描引擎列入黑名单，其中包括 Google、McAfee、

Yandex、Norton、ESET 和 AVAST 引擎。我们观察到，“软件”免费内容网站的 URL 被列入黑名单的比例显著高于其他类别，达到 23.12%，而其他所有类别的列入黑名单比例均不超过 12%。造成这一现象的原因之一是，这些网站经常通过使用不同的顶级域名（TLD）来频繁更换其域名。

要点： 为了评估免费内容网站相关的风险，我们利用 *VirusTotal* 和 *Sucuri* API，分析了这两类服务的网站域名和文件的恶意性。我们的分析发现，免费内容网站存在令人担忧的趋势，包括：(i) 免费内容网站更可能在域名级别与恶意性相关（占免费内容网站的 38%）；(ii) 它们更可能在文件级别与恶意性相关（占 45%）。这些趋势不仅体现在恶意性上——这导致大量网站被列入黑名单——还包括可被利用的漏洞，可能使访客遭受泄露攻击。我们的分析还发现，17% 的免费内容网站易受攻击。



(a) 书籍 (b) 游戏 (c) 电影 (d) 音乐 (e) 软件

图 9：无风险网站与高风险网站的决策边界。高风险网站是指具有潜在恶意意图的网站。请注意，这种恶意行为可以使用支持向量机进行特征描述（即确定）。

5.2. 风险建模

尽管我们迄今提供的见解颇为引人入胜，但仍有一个关键问题尚待解答：这些见解在多大程度上可以推广到同一人群和类型的网站，以用于风险评估？为回答这一问题，我们报告了利用简单的机器学习算法识别高风险网站的工作。

表 3：网站特征描述的说明。这些特征提取自三个来源：(i) 网站内容，(ii) 网站公开信息，(iii) 网站 SSL 证书信息。我们纳入了从 VirusTotal 和 Sucuri API 提取的特征，用于风险特征描述和潜在检测（D）。Org：特征的来源，*[c]*：分类特征，*[b]*：布尔特征（T/F），*[n]*：数值特征，*[p]*：百分比特征。

#	§	Type	描述	#	§	Type	描述
1	§4.1	[c]	网站使用的顶级域名名称	14	§4.2	[p]	网站中重定向内容所占的百分比
2	§4.1	[b]	未匹配的域名	15	§5.1	[b]	VirusToal API 检测到的域名
3	§4.1	[b]	已过期的 SSL 证书	16	§5.1	[b]	网站已检测到为恶意内容。
4	§4.1	[b]	SSL 证书无法验证	17	§5.1	[b]	包含恶意软件的网站
5	§4.2	[n]	HTTP 请求的平均数量	18	§5.1	[b]	已检测到网站存在网络钓鱼行为
6	§4.2	[n]	平均内容大小	19	§5.1	[b]	检测为恶意的文件
7	§4.2	[n]	页面加载时间	20	§5.1	[b]	检测为恶意的网址
8	§4.2	[p]	网站中图像的百分比	21	§5.1	[b]	被扫描引擎列入黑名单
9	§4.2	[p]	网站中脚本文件的百分比	22	§5.1	[b]	网站中的漏洞
10	§4.2	[p]	网站中字体内容的百分比	23	§5.2	[n]	网站 IP 地址生命周期
11	§4.2	[p]	网站中 HTML 文件的百分比	24	§5.2	[b]	使用/曾使用 Cloudflare 作为内容分发网络
12	§4.2	[p]	网站中 CSS 文件的百分比	25	§5.2	[b]	使用/曾使用 Akamai 作为内容分发网络
13	§4.2	[p]	网站中 XHR 内容所占的百分比				

高风险网站。 在我们的分析中，如果某个网站与以下任意情况相关联，则将其视为高风险网站：

(1) 恶意域名。 与负责恶意活动的URL相关联的网站被视为存在风险 [54]。

(2) 恶意文件。访问网站时，主机上会执行多个脚本。因此，我们认为任何包含恶意文件 [55], 的网站，无论其 $VirusTotal$ 标签如何，都是有风险的网站。

(3) 列入黑名单的*URL*。 列入黑名单可能是由于 (i) 大规模用户举报，或 (ii) 网站先前存在恶意行为（例如诈骗攻击）。因此，我们将所有列入黑名单的网站视为存在风险 [54]。

(4) 易受攻击的网站。 被$Sucuri$识别为易受攻击的网站，因其可能具有可利用性，因此被视为存在风险。

我们注意到，风险建模并不局限于免费内容网站。我们还将具有一个或多个上述方面的任何免费内容和高级网站视为

风险网站，否则视为无风险网站。

网站特征。 为了对与每项服务相关的风险进行建模，并遵循相关文献中的常见做法 [34, 54, 56, 57], ，我们利用前述提取的特征作为表示。具体而言，表 3 展示了我们用于表示每项在线服务的潜在特征全集，包括 SSL 证书、页面大小、加载时间、顶级域名和网站内容特征。此外，我们还使用 $SecurityTrails$ [58] 提取了另外三个特征：(i) 服务 IP 地址的生命周期；(ii) 网站是否正在使用或曾经使用 Cloudflare 作为内容分发网络（CDN）；以及 (iii) 网站是否正在使用或曾经使用 Akamai Tech 作为 CDN。

留出数据。 表 3 中由 $VirusTotal$ 和$Sucuri$ 获取的数据（#15–#25）被留出，仅用于模型验证。这样，我们便可以利用易于获取的网站质量指标对网站风险进行建模，而无需访问第三方信息。我们认为，与第三方风险数据相比，我们的轻量级建模会更加实用，因为第三方标签是基于报告以及在一段时间内积累的昂贵分析确定的。仅依赖第三方工具（例如 $VirusTotal$ ）来识别风险，会排除大量网站，包括那些新创建的、提供免费内容的网站。

风险边界。 考虑到上述特征，我们将高风险与低风险网站之间的边界可视化，如图 9 所示。具体而言，我们使用 t 分布随机邻域嵌入（t-SNE）可视化技术 [59]绘制网站的特征。随后，我们使用支持向量机估算风险边界，如图 9 中红色阴影区域所示。根据验证结果，我们发现，风险最高的网站聚集在一起，因为它们具有不同的网站特征。尽管存在一些局限性（例如，所采样的网站中可能出现假阳性），我们的模型仍能够以 86.81% 的准确率识别高风险网站。

要点： 我们针对免费内容网站的轻量级风险建模需求，采用由 17 个通用特征和文件相关特征构成的表示方法。结果表明，我们的建模方法效果显著，准确率达到 86.81%。

6. 结论与未来工作

免费内容网站是当今网络构成中一个有趣的组成部分，但与提供相同内容的其他网站相比，它们的特征尚未得到严格分析和理解。本文首次从多个域名和内容级维度以及风险概况出发，对此类网站进行了比较分析。我们整理的数据集为探索这一尚未充分研究的领域提供了宝贵资源，而我们的研究结果揭示了免费内容网站与高级网站之间的根本差异。

我们认为，本文的分析仅仅“触及了”这一重要问题的表面，仍有待进一步探索和采取行动。例如，我们的域名和内容级分析仅关注易于获取的元数据特征，未考虑更深入的特征，例如语言学特征、网络拓扑信息、区域信息、深层内容类型以及组织属性（例如，在 SSL 证书的情况下，还包括签名机构和托管基础设施）。所有这些维度都能够进一步揭示此类网站的特征，并构成我们未来的工作方向。最后，我们注意到，本分析利用的只是这些网站的单个快照，未考虑其特征的时间维度；这是一个非常有趣但也颇具挑战性的探索方向。

参考文献

[1] F. Hecker，《开设店铺：开源软件的商业运作》，IEEE《软件》，16（1）（1999），45–51。

[2] K. Greenhill, C. Wiebrands，无需图书馆：在线内容共享轻松免费的天地，VALA 2012：赋能电子未来。

[3] M. Carvajal, J. A. Garc´ıa-Avil´es, J. L. Gonz´alez，《众筹与非营利媒体：公共利益新闻新模式的出现》，《新闻实践》6（5-6）（2012）638–647。

[4] R. Snijder，《免费书籍的收益：一项衡量开放获取出版影响的实验》，《学习出版》23（4）（2010）293–301。

[5] J. R. Mayer，J. C. Mitchell，《第三方网络跟踪：政策与技术》，载于：IEEE安全与隐私研讨会，SP 2012，2012年5月21日至23日，美国加利福尼亚州旧金山，IEEE计算机学会，2012年，第413–427页。URL https://doi.org/10.1109/SP.2012.47

[6] C.-J.Liang，H.-J. Chen，《网站质量对客户关系绩效影响的研究》，《全面质量管理》20 (9) (2009) 971–988。

[7] T.-C. Lin，J. S.-C. Hsu，H.-C. Chen，《客户为在线音乐支付的意愿：免费心态的作用》，《电子商务研究杂志》14 (4)。

[8] J. Estrada-Jim´enez，J. Parra-Arnau，A. Rodr´ıguez-Hoyos，J. Forn´e，《在线广告：隐私威胁与保护方法分析》，Com- put. Commun. 100 (2017年) 32–51。网址：https://doi.org/10.1016/j.comcom.2016.12.016

[9] A. Alabduljabbar，R. Ma，S. Alshamrani，R. Jang，S. Chen，D. Mohaisen，《海报：免费内容网站风险的测量与评估》，载于：网络与分布式系统安全研讨会（NDSS），2022年。

[10] A.Alabduljabbar、R.Ma、S. Choi、R. Jang、S. Chen、D. Mohaisen，通过分析免费内容网站的 SSL

证书，了解免费内容网站的安全性：一项比较研究，载于：第一届网络安全与社会科学研讨会（CySSS），2022 年，第 19–25 页。

[11] A. Alabduljabbar、D. Mohaisen，通过自动化隐私政策分析与标注衡量免费内容网站的隐私维度，载于：《万维网大会》伴随会议，ACM，2022年，第860–867页。

[12] Z. Li、K. Zhang、Y. Xie、F. Yu、X. Wang，了解你的敌人：理解并检测恶意网络广告，载于：2012年ACM计算机与通信安全会议论文集，2012年，第674–686页。

[13] W. D. Groef、D. Devriese、N. Nikiforakis、F. Piessens，Flowfox：一款具有灵活且精确的信息流控制功能的网页浏览器，载于：ACM计算机与通信安全会议，CCS，ACM，2012年，第748–759页。URL https://doi.org/10.1145/2382196.2382275

[14] A. Nappa、R. Johnson、L. Bilge、J. Caballero、T. Dumitras，《克隆的攻击：共享代码对漏洞修补影响的研究》，载于：IEEE安全与隐私研讨会（SP），2015年，第692–708页。URL https://doi.org/10.1109/SP.2015.48

[15] VirusTotal，自动分析可疑文件和URL以检测恶意软件类型（2022年5月）。URL https://www.virustotal.com/

[16] Sucuri，网站安全检查与恶意软件扫描器（2022年5月）。URL https://sucuri.net/

[17] D. Perdices、J. E. L. de Vergara、I. Gonz´alez、L. de Pedro，深度学习时代的网页浏览隐私：超越虚拟专用网络与加密，《计算机网络》220 (2023) 109471。doi:10.1016/j.comnet.2022.109471。URL https://doi.org/10.1016/j.comnet.2022.109471

[18] H. Zou、J. Su、Z. Wei、S. Chen、B. Zhao，采用布朗距离协方差的高效跨域少样本网站指纹攻击，《计算机网络》219 (2022) 109461。doi:10.1016/j.comnet.2022.109461。URL https://doi.org/10.1016/j.comnet.2022.109461

[19] G. Sun、Z. Zhang、Y. Cheng、T. Chai，基于自适应分段网页文本的恶意网站检测，《计算机网络》216（2022）109236。doi:10.1016/j.comnet.2022.109236。网址 https://doi.org/10.1016/j.comnet.2022.109236

[20] A. Faroughi, A. Morichetta, L. Vassio, F. V. D. de Figueiredo, M. Mellia, R. Javidan，使用基于图的半监督学习进行网站域名分类，计算机网络 188 (2021) 107865。doi:10. 1016/j.comnet.2021.107865。URL https://doi.org/10.1016/j.comnet.2021.107865

[21] T. Chung, Y. Liu, D. R. Choffnes, D. Levin, B. M. Maggs, A. Mislove, C. Wilson，《测量并应用无效 SSL 证书：沉默的大多数》，载于：ACM 互联网测量大会，IMC，2016 年，第 527–541 页。

[22] T. Libert，揭示隐藏网络：对 100 万个网站上的第三方 HTTP 请求进行分析，CoRR abs/1511.00619。[23] A. Lavrenovs，F. J. R. Melon，排名前 100 万的网站的 HTTP 安全标头分析，载于：国际网络冲突会议（CyCon），2018 年，第 345–370 页。[24] I. Alsmadi，F. Mira，网站安全分析：检测方法和决策的变化，载于：2018 年第 21 届沙特计算机学会全国计算机会议（NCC），2018 年，第 1–5 页。[25] G. Tan，P. Zhang，Q. Liu，X. Liu，C. Zhu，F. Dou，自适应恶意 URL 检测：在存在概念漂移的情况下进行学习，载于：第 17 届 IEEE 计算、通信与网络安全和隐私国际会议以及第 12 届 IEEE 大数据科学与工程国际会议，

TrustCom/BigDataSE 2018，美国纽约州纽约，2018 年 8 月 1–3 日，IEEE，2018 年，第 737–743 页。

[26] R. Masri、M. Aldwairi，使用VirusTotal、URLVoid和TrendMicro进行自动恶意广告检测，载于：信息与通信系统国际会议（ICICS），2017年，第336–341页。[27] K. Patil，隔离浏览器扩展的恶意内容脚本，《信息隐私与安全集成国际期刊》3（1）（2017年）18–37。

[28] V. R. L. Shen、C. Wei、T. T. Juang，使用高级模糊Petri网进行JavaScript恶意软件检测，载于：2018年国际机器学习与控制论会议（ICMLC 2018），中国成都，2018年7月15–18日，IEEE，2018年，第511–514页。

[29] R. Wang、Y. Zhu、J. Tan、B. Zhou，基于混合分析的恶意网页检测，《信息安全与应用杂志》35（2017年）68–74。[30] C. Johnson、B. Khadka、R. B. Basnet、T. Doleck，利用深度学习检测和分类恶意URL，《无线移动网络、普适计算与可信应用杂志》11（4）（2020年）31–48。[31] A. Desai、J. Jatakia、R. Naik、N. Raul，使用机器学习进行恶意网页内容检测，载于：2017年第2届IEEE电子、信息与通信技术最新趋势国际会议（RTEICT），IEEE，2017年，第1432–1436页。[32] J. M. IV、D. Bhansali、M. Gratian、M. Cukier，HTTP头部特征检测恶意网站的综合评估，载于：欧洲可信计算会议（EDCC），2019年，第75–82页。[33] Z. Li、K. Zhang、Y. Xie、F. Yu、X. Wang，了解你的敌人：理解和检测恶意网络广告，载于：ACM计算机与通信安全会议（CCS），2012年，第674–686页。

[34] A. Mohaisen、O. Alrawi，揭示Zeus：恶意软件样本的自动分类，载于：万维网（WWW），2013年，第829–832页。URL https://doi.org/10.1145/2487788.2488056

[35] H. Kang、J. Jang、A. Mohaisen、H. K. Kim，结合创建者信息，使用静态分析检测并分类安卓恶意软件，《国际分布式传感器网络期刊》11。网址：https://doi.org/10.1155/2015/479174[36] H. Kang、J.-w. Jang、A. Mohaisen、H. K. Kim，结合创建者信息，使用静态分析检测并分类安卓恶意软件，《国际分布式传感器网络期刊》11 (6) (2015) 479174。[37] A. Mohaisen、O. Alrawi、M. Mohaisen，AMAL：高保真、基于行为的自动化恶意软件分析与分类，《计算机安全》52 (2015) 251–266。网址：https://doi.org/10.1016/j.cose.2015.04.001[38] H. Alasmary、A. Khormali、A. Anwar、J. Park、J. Choi、A. Abusnaina、A. Awad、D. Nyang、A. Mohaisen，分析并检测新兴物联网恶意软件：一种基于图的方法，《IEEE物联网期刊》6 (5)。[39] R. Masri、M. Aldwairi，使用VirusTotal、URLVoid和TrendMicro自动检测恶意广告，载于：2017年第8届信息与通信系统国际会议（ICICS），2017年，第336–341页。[40] A. Geniola、M. Antikainen、T. Aura，下载门户和免费软件安装程序的大规模分析，载于：《安全 IT 系统——第22届北欧会议 NordSec 2017》，爱沙尼亚塔尔图，2017年11月8日至10日，会议论文集，《计算机科学讲义》第10674卷，Springer，2017年，第209–225页。[41] R. Rivera、P. Kotzias、A. Sudhodanan、J. Caballero，高成本免费软件：一种

对下载门户滥用行为的系统分析，《IET信息安全》13 (1) (2019) 27–35。[42] A. S. Manjeri、R. Kaushik、M. Ajay、P. C. Nair，一种利用URL特征检测恶意网站的机器学习方法，载于：2019年第3届电子、通信与航空技术国际会议（ICECA），2019年，第555–561页。[43] A. K. Singh、N. Goyal，机器学习属性在检测恶意网站方面的比较，载于：第11届通信系统与网络国际会议（COMSNETS 2019），印度班加罗尔，2019年1月7日至11日，IEEE，2019年，第352–358页。[44] D. Liu、J. Lee，通过使多个网络垃圾信息无效化，基于卷积神经网络检测恶意网站，《IEEE Access》8 (2020) 97258–97266。[45] PyWebCopy，Pywebcopy：使用Python抓取并保存网页和网站的工具（2022年5月）。网址：https://pypi.org/project/pywebcopy/[46] D. state，域名工具、统计数据、新闻、论坛和目录（2022年5月）。网址：https://www.domainstate.com/registrar-tld-breakup.html[47] Google，面向Chrome和Firefox的基于用户举报的Google安全浏览（2022年5月）。网址：https://safebrowsing.google.com/safebrowsing/report_general/[48] H. I. Archive，前1,000,000名：页面重量报告（2022年5月）。网址：https://httparchive.org/reports/page-weight?lens=top1m[49] Pingdom，网站性能与可用性监控（2022年5月）。网址：https://www.pingdom.com/[50] A. Cohen、N. Nissim、Y. Elovici，Maljpeg：一种基于机器学习的恶意JPEG图像检测方案，《IEEE Access》8 (2020) 19997–20011。

[51] L. South、D. Saffo、M. A. Borkin，社交媒体中诱发癫痫发作的 GIF 检测与防御，载于：CHI '21：计算机系统中的人因大会，虚拟活动/日本横滨，2021 年 5 月 8—13 日，ACM，2021 年，第 273:1–273:17 页。

[52] W. Yost、C. Jaiswal，Malfire：用于恶意内容检测与防护的恶意软件防火墙，载于：第八届 IEEE 年度普适计算、电子与移动通信大会，UEMCON 2017，美国纽约州纽约市，2017 年 10 月 19—21 日，IEEE，2017 年，第 428—433 页。

[53] D. Jung、S. Lee、I. Euom，Imagedetox：隐藏于图像文件中的恶意代码的中和方法，Symmetry 12（10）（2020）1621。

[54] A. G. West、A. Mohaisen，《恶意软件中出现的网络端点的基于元数据的威胁分类》，载于：DIMVA 会议论文集，2014 年。[55] A. E. Kosba、A. Mohaisen、A. G. West、T. Tonn、H. K. Kim，《ADAM：恶意网页的自动检测与归因》，载于：第 15 届信息安全应用国际研讨会（WISA），2014 年，第 3–16 页。[56] A. Mohaisen、O. Alrawi，《AMAL：高保真、基于行为的自动化恶意软件分析与分类》，载于：WISA 会议论文集，2014 年。[57] A. Mohaisen，《面向恶意网页内容的自动化轻量级检测与分类》，载于：IEEE HotWeb 会议论文集，2015 年。[58] SecurityTrails，《探索任何互联网资产的完整当前数据和历史数据》。IP 与 DNS 历史记录（2022 年 5 月）。网址：https://securitytrails.com/[59] L. Van der Maaten、G. Hinton，《使用 t-SNE 可视化数据》，《机器学习研究期刊》 9 （ 11 ）。