八卦：从邮件列表讨论中自动识别恶意域

中国四川大学城隍
opcodesec@gmail.com

双浩
美国加州大学圣巴巴拉分校
shuanghao@cs.ucsb.edu

卢卡·因韦尔尼齐
美国加州大学圣巴巴拉分校
invernizzi@cs.ucsb.edu

刘佳勇
中国四川大学
jylscu@gmail.com

永方
中国四川大学永夫
angscu@gmail.com

Christopher Kruegel 美国加州大学圣巴巴拉分校和 Lastline, Inc.
chris@cs.ucsb.edu

Giovanni Vigna 美国加州大学圣巴巴拉分校 & Lastline, Inc.
vigna@cs.ucsb.edu

摘要

域名在网络犯罪中起着至关重要的作用，因为它们可以识别提供恶意内容（如恶意软件、特洛伊木马二进制文件或恶意脚本）的主机，作为命令和控制服务器运行，或在恶意网络基础设施中扮演其他角色.为防范互联网攻击和诈骗，运营商广泛使用黑名单检测和阻断恶意域名和IP地址。现有的黑名单通常是通过爬取可疑域、手动或自动分析恶意软件以及从蜜罐和入侵检测系统收集信息而生成的。不幸的是，这样的黑名单很难维护，而且通常对新的攻击反应很慢。

安全专家建立并加入邮件列表来讨论和共享情报信息，这为识别新出现的恶意活动提供了更好的机会。在本文中，我们设计了 Gossip，这是一种通过使用自然语言处理和机器学习技术，基于技术邮件列表（特别是安全相关主题）中的讨论分析自动检测恶意域的新方法。我们确定了一组从电子邮件线程、参与讨论的用户和内容关键字中提取的有效特征，以从邮件列表中推断出恶意域，而无需实际抓取可疑网站。我们的结果表明，八卦实现了很高的检测精度。此外，我们系统的检测通常比现有的公共黑名单早几天或几周。

ASIA CCS '17, 2017 年 4 月 2 日至 6 日，阿拉伯联合酋长国阿布扎比

二：http://dx.doi.org/10.1145/3052973.3053017

关键字

恶意软件检测；自然语言处理；黑名单

1. 介绍

因特网发展迅速，提供范围广泛的服务。然而，它也已成为网络犯罪分子的诱人目标，他们会执行恶意活动，例如垃圾邮件广告、金融欺诈和路过式下载攻击。这些攻击通常与域名相关联，指向包含恶意内容或作为命令和控制服务器运行的主机。当服务器被识别为恶意服务器时，执法和安全专家将其关闭，但不法分子不断注册新域以支持非法活动。安全公司和研究人员开发了许多黑名单，例如 Spamhaus [5] 或 PhishTank [4]，以检测和过滤攻击中使用的域。现有系统通过分析网络流量、爬取可疑域或部署蜜罐来识别非法活动来生成黑名单。尽管黑名单是检测潜在攻击的最广泛使用的方法之一， 它存在两个主要问题：覆盖范围有限和检测延迟。最近的研究表明，传统的黑名单只捕获了一小部分恶意域 [30, 35]，并且在攻击发生后它们在捕获恶意域方面经历了相当大的延迟 [39, 43]。

在攻击的早期阶段（即在黑名单成功报告恶意域之前），网络管理员或安全专家可能已经注意到异常活动，并且他们倾向于找到有效的方法来确认他们的观察和交流发现。 Web 论坛或邮件列表是讨论和共享此类信息的理想场所。用户发布问题域后，其他人可以回复并提供额外的证据，这有助于确定相关域是否确实是恶意的。这种在线讨论有几个优点。 (1) 综合多方面的观察得出结论。 (2) 相关人员的知识和经验

参与讨论的安全专业人士有助于准确识别潜在的攻击。 (3) 由于这些讨论主要集中在早期出现的恶意活动，因此只有少数受害者受到攻击。如果能够构建一个使用这些信息并及时生成黑名单的检测系统，将有助于防止攻击的更大规模传播。然而，邮件列表中发布的内容是非结构化的，需要人工来发现有用的信息。迄今为止，很少有人进行自动识别邮件列表中讨论的威胁的研究，尽管语言分析已成功应用于其他与安全相关的问题，例如使用文体学来识别匿名博主 [37]。

在这项工作中，我们设计并实现了 Gossip，这是一种检测系统，可以自动从邮件列表的讨论中提取恶意域。主要的直觉是，安全专家之间的这些讨论有助于确定涉及非法活动的域名的特征。八卦依赖于语言模式、电子邮件线程信息和参与讨论的用户的特征。这些特征可以很容易地以很小的开销提取（与网站爬行或恶意软件分析相比）。我们使用机器学习技术构建统计分类模型，并评估 40 多个邮件列表的性能。我们的结果表明，八卦实现了很高的检测精度。此外，Gossip 的检测通常比现有的公共黑名单早几天或几周；特别是，40% 的域可以在出现在黑名单前 10 天被检测到。我们手动验证预测结果， 并找到数百个其他黑名单无法识别的恶意域。网络管理员可以使用我们系统的输出来查找其网络中的攻击流量。Gossip 可以有效地补充其他检测方法并提高对恶意域的防御。

总之，我们做出以下贡献：

- • 我们开发了 Gossip，这是一种新颖的检测系统，它使用语言和机器学习技术根据邮件列表讨论自动识别恶意域。

• 我们确定了四个新功能系列，以利用安全专家对可疑域的讨论和意见。 (1) URL 特征捕获嵌入 URL 的模式； (2) 线程特征基于与线程中的多个电子邮件关联的元数据； (3) 参与者特征关注参与电子邮件线程讨论的参与者； (4) 上下文特征检查所分析域周围的文本。 Gossip 关注邮件列表中的数据，并且需要很少的辅助信息来导出这些特征。

- •我们在现实世界的邮件列表上评估八卦。地面实况信息是从公共黑名单和人工检查中收集的。我们表明 Gossip 能够准确识别恶意域，导致 94% 的检测率和零误报率。我们的结果表明，与现有的黑名单相比，八卦提供了更早的检测。

2. 系统设计

在本节中，我们将介绍系统的设计和内部细节。Gossip 旨在从讨论安全相关主题的邮件列表中快速准确地找到恶意域。我们面临两个主要挑战。首先，邮件列表中的数据和信息没有固定的格式，内容文本也没有结构化。其次，很难识别表明恶意域的有效特征。为了实现我们的目标，我们开发组件来解析原始数据，提取一组显着特征，并构建一个模型来对域是否涉及恶意活动进行分类。图 1 显示了我们系统的架构。输入是来自邮件列表的电子邮件消息。值得注意的是，Gossip 可以轻松扩展以处理从其他平台（例如新闻组或网络论坛）收集的消息。输出是具有分类分数的域列表，其中较高的分数表示域更有可能是恶意的。 Gossip 由四个组件组成：预处理器、域提取器、特征解析器和自动分类器。接下来我们将更详细地描述每个组件的功能和实现。

2.1 预处理

作为第一步，我们需要处理原始输入数据。来自邮件列表的消息以 mbox 格式（RFC 4155）存储在单个文件中。我们开发了一个电子邮件解析器，将邮箱内容分离为 MIME 格式（RFC 5322）的单独电子邮件消息，并提取消息属性，包括标头信息和消息正文。特别是，四个标题字段有助于稍后的特征提取（见第 3 节）：“From”字段表示消息的作者，“Organization”字段表示作者所属的公司或组织，“Content-Disposition”字段显示电子邮件是否有附件，“Subject”字段包含表示电子邮件主题的词。“日期”字段显示消息发送的时间，这使我们可以将邮件列表中的讨论时间与域最终包含在其他黑名单中的时间进行比较（参见第 4.2.2 节）。

邮件正文不仅包含干净的文本，还可以包含 HTML 元素或特殊编码的字符。文本清理过程从呈现的 HTML 中提取纯文本并删除特殊字符。我们丢弃使用加密（例如 PGP 加密）的消息，因为消息内容不可访问。此外，我们的检测信息很少的文本部分（例如用户签名）被排除在进一步分析之外。为简单起见，我们过滤掉非英语语言的电子邮件。

当一封电子邮件发布在邮件列表中时，其他人可以回复并参与讨论，从而形成一个会话线程。我们开发了一个模块，该模块执行线程分组以关联属于同一对话的电子邮件。每封电子邮件都有一个唯一的“Message-ID”。回复消息具有“In-Reply-To”字段，指的是它回复的上一条消息的“Message-ID”（线程中的第一封电子邮件具有空的“In-Reply-To”字段）。在这个过程之后，我们根据标题字段“In-Reply-To”和“Message-ID”链接同一线程中的电子邮件。预处理器组件的输出是属性

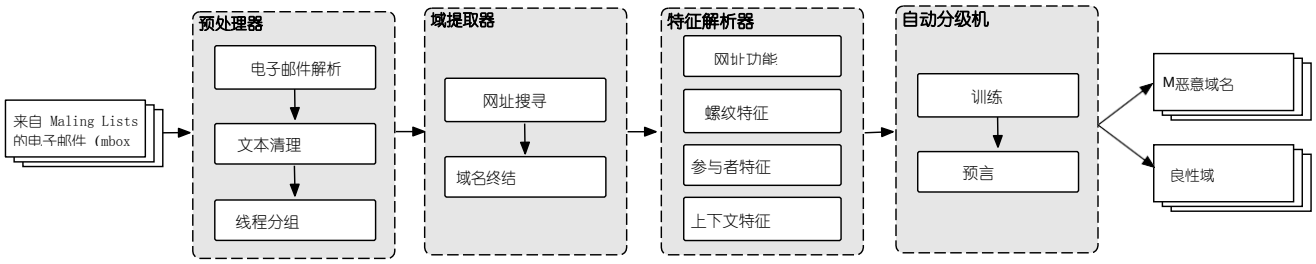

图 1: Gossip 的架构。该系统包含四个组件（具有各种内部模块）。

电子邮件消息（包括标题和正文）和相关线程。

2.2 域提取器

我们的分析目标是邮件列表中讨论的域，它们对应于部分 URL 字符串。我们先在消息体中找到URL，然后推导出域名。虽然乍一看从电子邮件文本中提取 URL 似乎微不足道，但用户采用的任意约定或变体通常会变形标准 URL 格式并带来额外的困难。由于可疑 URL 很可能会导致指向托管恶意内容的网站，因此安全专业人员倾向于在 URL 中替换或插入字符以避免意外点击。例如，一个 URL http://www.malwareexample.com/index.html 可以混淆为 hxxp//www[.]malwareexample[.]com/index.html，使其无法直接点击。我们手动构建正则表达式签名来匹配模式，并开发一个 URL 搜索模块来派生正确的 URL。

我们随后解析域最终确定模块中的 URL 以获得候选域。在我们的分析中，我们主要关注二级域名。对于前

充足，来自上面的 URL http://www.malwareexample。 com/index.html，提取的域是malwareexample。 com。一些国家代码二级域，例如 .com.cn 和 .co.uk，向公众提供子域注册（如 Mozilla 公共后缀列表 [3] 中所定义）。对于这些域，我们提取三级域名作为分析目标。一个域可以出现在多个线程中，因此我们将每个域与所有提到它的电子邮件线程相关联以进行进一步处理。

2.3 特征解析器

在前面的步骤之后，每个域都与一组讨论线程和相应的电子邮件消息相关联。一个关键问题是：什么特征最能表征恶意域？我们探索邮件列表中讨论行为和内容的属性。我们将特征分为四类。 URL 特征检查 URL 的词法模式和有关域的辅助信息，例如 URL 是否按标准方案键入或域是否在 Alexa 顶级列表中。线程特性侧重于线程的属性。我们描述了域的讨论方式，例如有多少电子邮件参与了讨论或线程是否包含哈希值（通常用于指纹恶意软件）。参与者特征调查谁参与讨论。在有人发布有关可疑域的问题后，其他人会回复消息以提供答案或其他观察结果。参与者的声誉表明人们可以在多大程度上信任讨论并得出该领域存在问题的结论。上下文特征考虑讨论内容的语义模式，特别是基于围绕分析域的文本。值得注意的是，Gossip 仅使用从讨论线程派生的功能，不需要主动扫描可疑域或 URL。在第 3 节中，我们更详细地解释了这些特征，并讨论了如何表示它们。

为了避免由不同的特征尺度引起的偏差，我们对特征应用了归一化过程。对于每个特征值 fn Fn，归一化值按照等式 1 计算，其中 max(Fn) 和 min(Fn) 分别对应于 Fn 中的最大值和最小值。min-max归一化将[0, 1]范围内的特征值进行转换。

$$归一化(f_n) = \frac{f_n - 最小值(Fn)}{最大值(F) - 最小值(F)} \quad (1)$$

2.4 自动分级机

我们将这些特征整合到分类方法中，以确定讨论中的域是否倾向于参与恶意活动。在我们的应用场景中，特征矩阵很大（大约 20,000 个特征）并且稀疏（有很多零值）。我们在实验中将包括朴素贝叶斯、支持向量机 (SVM) 和随机森林 [13] 在内的多种分类方法与真实数据进行了比较。特别是随机森林，它结合了一组决策树来投票选出最终结果，在我们的实验中产生了最好的准确性和运行性能。因此，我们选择使用随机森林作为 Gossip 中的自动分类器。在训练阶段，我们使用标记的基本事实（基于人工检查和现有黑名单，见第 4 节）来构建分类模型。在预测阶段，分类器为每个域输出 [0, 1] 范围内的分数， 其中较高的分数表示该域更有可能是恶意的。运营商可以设置适当的阈值来实现误报率和检测率之间的权衡。如果域的分类分数大于预定义的阈值，则 Gossip 将该域报告为恶意域。

3. 特征

对于每个正在分析的域，我们首先汇总所有提及它的电子邮件线程。我们从这些电子邮件中提取特征向量，然后将其输入到我们的机器学习组件中。在本节中，我们将详细介绍每个功能类别。

3.1 网址功能

我们从每个 URL 中提取两个特征。第一个是流行度特征，它指示 URL 的域名是否包含在 Alexa 前 100 万个域中。此功能引入了关于我们分类的先验知识，为此我们倾向于在查看电子邮件线程的内容之前将非常流行的域分类为良性域。

第二个特征表示该URL是否可能导致二进制下载，通过检查该URL是否包含“.exe”扩展名来实现。导致二进制下载的 URL 更有可能是恶意的。

3.2 螺纹特征

我们识别所有提及每个 URL 的电子邮件线程，并从线程集合中提取特征。

电子邮件主题中使用的术语。电子邮件的主题通常提供电子邮件内容的简明概述。因此，在去除常用词（例如，停止词、“Re:”、.）和标点符号后，我们构建了一个表示这些主题中使用的术语的词袋二元向量。例如，“邮件列表成员提醒”一词的存在表明内容是良性的。电子邮件主题是从邮件标题中的“主题”字段中提取的。

回复数。我们在单个功能中指出提到 URL 的线程的总长度。出现频率高的 URL 通常是粒子签名的一部分，这表明它是良性的。另一方面，频率较低的 URL 更有可能是恶意的。正如我们所讨论的，回复数是根据“Message-ID”和“In-Reply-To”字段计算的。

域的数量。我们计算与每个 URL 同时出现的不同域的数量。此功能表明存在“URL 转储”，其中一些电子邮件包含已被同一攻击者使用或受害的域列表。

IP 地址数。与之前的功能类似，我们计算与每个 URL 同时出现的 IPv4 地址总数。长长的 IP 地址列表通常用于识别受攻击者控制的受感染主机。我们解析消息内容并提取四点格式的字符串（每个整数范围从 0 到 255）作为 IP 地址。

附件数量。在讨论特定的恶意软件时，一些参与者将附加证据作为附件。这些附件包括恶意软件样本、网络转储或受影响用户的列表。我们包括一个计算它们的功能。附件由每封电子邮件中的“Content-Disposition”标头字段指示。

组织数量。一些公司将其 SMTP 服务器配置为包含非标准电子邮件标头所有外发邮件中的“组织”。该字段有时以复数形式和/或英式变体拼写。我们将组织名称编码为二进制向量，并将其添加到我们的特征中。特别是，我们为每封电子邮件生成一个单热向量，并将它们全部聚合到一个向量中，对于参与讨论的每个组织，该向量都将具有非零值。

已知恶意软件检测服务的数量。我们计算 URL 的数量，与正在分析的 URL 位于同一位置，指向已知的恶意软件检测服务（例如，VirusTotal [20]）。为此，我们参考了我们手动编译的这些服务的简短列表。

加密摘要的数量。我们计算与所分析的每个 URL 位于同一位置的 SHA1 和 MD5 摘要的数量，因为这些摘要通常用于唯一识别恶意软件样本。我们利用 SHA1 和 MD5 摘要分别对应 40 和 32 个十六进制数这一事实，并计算符合条件的字符串。

3.3 参与者特征

参与邮件列表的成员具有不同的经验、资源和技术背景。因此，他们对讨论的贡献可能会有所不同。某些人（尤其是安全专家）的响应更有可能表明存在攻击或恶意活动。我们利用用户的活动并检查谁参与了对话线程。参与者的身份可以通过他们的电子邮件地址轻松区分。

首先，我们从数据集中提取所有发件人的电子邮件地址，并使用 MD5 算法计算它们的哈希值，以避免泄露任何私人信息。其次，我们将这些字符串用作参与者特征：如果电子邮件地址显示回复线程，则该值将设置为二进制值 1。最后，我们生成一个映射矩阵以指示哪些成员参与了有关域的讨论。

3.4 上下文特征

在最后一组特征中，我们提取了被分析的每个 URL 周围文本的特征。一个挑战是如何找到关键字以准确指示 URL 的声誉。传统的基于整个文本或段落的词袋方法有一些局限性，例如包含许多与主题不密切相关的常用词或句子。以前的研究只侧重于分析 URL 本身的语言模式 [18, 48]。

我们开发了一种从上下文关键字生成特征的方法。启发式是基于这样一种直觉，即靠近所讨论的 URL 的邻居词可以代表人类分析师对 URL 的判断。

我们从文本的窗口部分构建词袋特征向量（最多 20 个词，最近的一项研究表明，英语句子通常包含少于 20 个词 [17, 21]），以网址。如果大量负面含义的词与域相关联，则该域更有可能是恶意的。例如，考虑以下来自真实数据的片段：

> 第一阶段的后门是新编译的，与 IE 0day 网络钓鱼浪潮中使用的相同版本略有修改，并且重新使用了一个 c2 域 xxxxx.com 是引人注目的。和

域与一些恶意活动相关联。

从文本描述中，人类用户可以推断出域“xxxxx.com”是恶意的。我们希望 Gossip 能够自动提取能够代表域上讨论意见的词。本段中的一系列关键字可以指示有关域的属性或活动信息。为了选择代表词来反映领域的声誉，我们首先去除靠近实体位置的停用词（经常使用的常用词，在句子中没有任何意义），选择一定区域来限制词数，并提取最能表明域恶意性的词。上述示例的输出结果如下所示：

~~第一阶段的后门是新编译的，与 IE 0day 网络钓鱼浪潮中使用的相同版本略有修改，并且重新使用了一个 e2 域 xxxxx.com 是引人注目的。并且该域与一些恶意活动相关联。~~

为了从电子邮件线程中自动提取上下文特征，我们使用四个步骤：文本清理、删除停用词、词干提取和提取上下文词。接下来我们详细介绍每个步骤。

文字清洗。除了纯文本（如 HTML 页面、非 ASCII 字符等）之外，还有多种电子邮件格式和编码方式。电子邮件标头中的“ContentType”表示消息的内容类型。编码方法可以在“Content-Transfer-Encoding”字段中找到。为了正确提取内容并删除不相关的 HTML 代码，我们使用 Beautiful Soup Python 库 [41] 解析 HTML 标签并对 Base64 编码的内容进行解码。此外，一些标点符号对于表达句子的意思和表明讨论的态度至关重要。例如，一些标点符号，如“！”、“？”，可能会影响句子的意思，而另一些标点符号，如“：”、“-”，则不代表用户的强烈情感。为了构建一个具有代表性的特征集，我们构建了一个标点符号列表并将这些标点符号包含到单词标记化中。

删除停用词。我们首先检查了 NLTK 库 [44] 中的停用词，但该库只包含有限的 127 个英语单词。谷歌研究团队计算并发布了 10,000 个最常用的英语单词（按频率排序）[19]。我们将 Google 列表中排名前 2000 的英文单词和 NLTK 库中的所有停用词组合在一起，然后手动删除具有情感意义的单词。最终，我们获得了 1,550 个停用词并将它们排除在我们的分析中。阻止。词干提取是获取其他时态格式动词（例如现在时、过去时或将来时）的词根形式的过程。在我们的工作中，我们将不同的时态格式词视为一个特征，并使用 Porter 词干提取模块 [49] 将每个词转换为其词根形式（这比其他词干提取算法 [22、40] 更不激进）。提取上下文词。要获取有关正在分析的域的上下文信息， 我们从提到域的所有线程中提取文本标记。

虽然天真的选择是选择线程中显示的所有标记，但该方法会使分类器混淆：用户可能会在同一个线程中讨论良性和恶意域。在线程级别使用词袋方法会导致包含与分析域无关甚至误导性的词标记。解决此问题的关键观察是，紧邻每个分析域的令牌最能表明该域的潜在恶意性。因此，对于每个域，我们只提取接近其在文本中位置的文本标记子集。

4. 评估

我们评估了 Gossip 在真实世界邮件列表数据上的表现。首先，我们描述了我们如何收集数据集并获得基本事实。然后，与现有黑名单相比，我们展示了关于检测精度和时间的实验结果。

4.1 数据集

我们使用的邮件列表数据包括两种类型的列表：专注于讨论恶意活动和攻击的安全邮件列表，以及讨论一般计算机相关问题的公共邮件列表。这些邮件列表提供了人们利用在线论坛共享信息和寻找解决方案的真实示例。我们收集了一些黑名单来标记域并使用它们来评估我们系统的性能。

安全邮件列表。来自不同组织或企业的安全专业人员建立邮件列表以交换意见和安全意见。由于此类邮件列表可能包含有关网络配置或公司安全产品的私人信息，因此公众通常无法访问内容，订阅请求需要版主批准。我们订阅了六个安全邮件列表并收集了从 2009 年 6 月到 2015 年 5 月的电子邮件讨论。这些邮件列表专注于与安全相关的主题，并发布有关可疑域的问题。处理原始数据后，我们提取了 10,780 个电子邮件线程和 2,649 个不同的域。

公共邮件列表。我们从公共邮件列表中收集数据，以进一步检查我们系统的稳健性。公共邮件列表提供了一个平台，人们可以在其中讨论与计算机相关的一般问题。讨论内容不太可能包含恶意域，因此我们使用公共邮件列表作为控制组，以确保 Gossip 在良性域上几乎没有错误分类（参见第 4.3 节）。可以在“MARC”存档（marc.info）中找到邮件列表的集合，该存档存储了来自 700 万用户的 3,500 多个邮件列表中的 7,000 万封电子邮件。从 1999 年 11 月到 2015 年 9 月，我们随机抽取了 43 个邮件列表并抓取了 76,380 个电子邮件线程，其中包含 3,898 个不同的域名。

黑名单标签。我们从公共黑名单和在线恶意软件检测服务中抓取数据以获取基本事实信息。值得注意的是，现有的黑名单并不能提供完美的 ground truth，因此我们进一步对 Gossip 的检测结果进行了人工验证（见 4.2.1 节）。一些黑名单服务有时间戳信息来指示域何时被列入黑名单，这使我们能够比较 de-

表 1：安全邮件列表中 244 个样本域的统计数据（作为训练数据）。

类别	计数和百分比
恶意域名	93 (38.11%)
良性域	141 (57.79%)
不确定的域	10 (4.10%)

保护现有的黑名单和我们的系统。请注意，在我们的实验中，我们仅收集并与公共黑名单进行比较。虽然商业黑名单包含更及时的信息，但 Gossip 有助于从人类分析师的在线讨论中自动提取威胁。我们使用的黑名单来自三个主要组。

• •多个黑名单的提要。我们获得了一个提要，其中收集了超过 75 个不同的公共黑名单 [29]，包括 malwaredomainlist [2]、abuse.ch [6]、MalcOde [1]、Spamhaus [5] 和 PhishTank [4]。收集时间跨度为2011年7月至2015年6月，数据涵盖涉及各种恶意活动的域名，例如网络钓鱼、垃圾邮件和偷渡式下载攻击。给定一个域名，提要会显示将该域报告为恶意域的特定黑名单以及相应的黑名单周期。• VirusTotal [20]。为了补充上述黑名单提要，我们查询了 VirusTotal，它集成了 60 多种不同的安全产品。然而，VirusTotal 会产生不可忽略的误报数 [45]，因此，只有当不止一种安全产品被标记为有问题时，我们实验中的域才会被标记为恶意域。VirusTotal 提供时间戳以指示首次分析某个域的时间。• McAfee SiteAdvisor [33]。另一个资源是 SiteAdvisor，这是 McAfee 提供的一项免费在线服务，用于检查某个域是否与恶意活动相关联。我们将从邮件列表中找到的域提交给 McAfee SiteAdvisor。尽管该服务返回许多恶意域标签，但不足之处是它不提供与检测相关的时间戳。

在我们的实验中，如果上述任何黑名单报告该域与非法活动相关，我们就会将该域视为恶意域。

4.2 安全邮件列表的实验

我们首先检查 Gossip 在安全邮件列表上的表现，并将结果与现有的黑名单进行比较。一个挑战是不确定安全邮件列表中讨论的哪些域是良性的。所以我们对一组域进行采样并手动检查它们是否是恶意的。为确保我们的样本具有代表性，我们不会从同一电子邮件线程（具有相似特征）中选择多个域。表1显示了抽样域的人工检查结果。我们使用 93 个手动检查的恶意域和 141 个良性域作为训练数据。

为了获得无偏见的测试数据，我们删除了出现在与训练数据相同的讨论线程中的域。换句话说，我们将训练数据和测试数据分开，使它们不相关。我们获得来自安全邮件列表的 2,405 个域作为测试数据集。

4.2.1 准确性和手动验证

我们在训练数据上使用随机森林算法学习分类模型，并将该模型应用到测试数据上。每个域的输出结果是一个分数，分数越高表示该域越有可能是恶意的。Gossip 归类为恶意的域取决于所选的分数阈值。在我们的实验中，我们设置了一个 0.4 的阈值（该值在 [0, 1] 范围内），这导致检测到大部分列入黑名单的域。有 1,559 个恶意域被 Gossip 归类为恶意域。黑名单报告了邮件列表中讨论的 693 个恶意域，其中 608 个域 (88.7%) 也被 Gossip 捕获。结果表明，Gossip 可以覆盖现有黑名单检测到的大部分恶意域。回想一下，我们没有关于恶意域和良性域的完美基本事实。为了进一步检验我们的结果有多准确，我们根据检测分数对域进行降序排序，并手动调查前 200 个域、分数刚好高于 0.4 阈值的 100 个域（排名 1,460-1,559）和最后 100 个域（即接近末尾的域）列表中的 2,405 个域）。表2显示了黑名单域的数量（第二列）和非黑名单域的人工检查结果（第三至第五列）。我们看到分类得分最高的域包含相当多的黑名单域：前 100 个域有 60 个被列入黑名单，接下来的 100 个域有 62 个被列入黑名单。当我们通过检查域的 Web 内容和邮件列表讨论中的证据来手动调查非黑名单域时，在 1-100 和 101-200 组中，我们发现 90.0% 和 76.4% 的非黑名单域名实际上是恶意域名，但被现有的黑名单遗漏了。 在排名1460-1559的100个域名中，还有大量的恶意域名没有被列入黑名单。观察表明，八卦补充了现有的黑名单，并捕获了更多的恶意域。在 5.1 节中，我们进一步分析了误报，表明白名单可以显着提高八卦的准确性。表中的最后一行显示了分类分数最低的 100 个域，它们应该是良性域。很少有出现在黑名单中或通过人工检查确定为恶意的。结果表明，我们的系统可以有效地区分恶意域和良性域。

4.2.2 检测时序比较

检测时间是黑名单和检测系统的重要指标，因为早期检测可以防止恶意活动传播给更多受害者。我们调查和比较恶意域何时出现在黑名单上以及何时可以被 Gossip 检测到。正如我们提到的，McAfee SiteAdvisor 不在数据中提供时间戳，因此我们将其排除在分析之外。有了上面的 0.4 阈值，我们有 476 个恶意域被信誉服务（不包括 McAfee SiteAdvisor）和 Gossip 标记。

我们使用讨论域的线程的最后一封电子邮件的时间戳（在“日期”标题字段中）作为时间

表 2：手动检查 Gossip 归类为恶意域的结果（按分类分数排名）。在每一行中，粗体数字的组合是恶意域的基本事实。

秩 地区	列入黑名单	非黑名单		
		恶意的	良性	不确定
1-100	**60**	**36** (90%)	4 (10%)	0(0%)
101-200	**62**	**29** (76%)	8 (21%)	1(3%)
...	...	.....		...
1,460-1,559	**20**	**22** (27%)	55 (69%)	3(4%)
...	...	.....		...

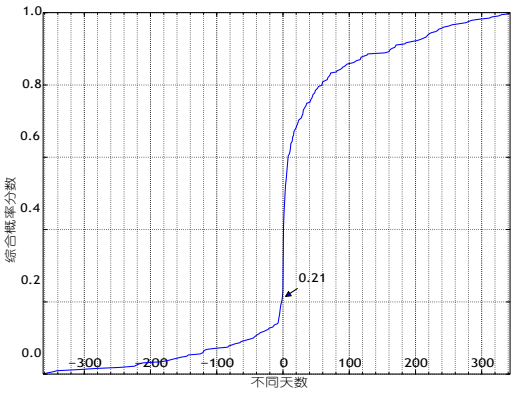

图 2：Gossip 和公共黑名单之间的检测时间差异分布（一年周期）。

当 Gossip 可以检测到域时。每个黑名单都提供第一次和最后一次将域标记为
恶意，形成黑名单窗口。由于不同的黑名单会产生多个黑名单窗口，我们需要找到合适的窗口与Gossip进行比较。计算时间差的详细过程如算法1所示。

图 2 显示了列入黑名单时间和 Gossip 进行检测之间的天数分布。我们只显示一年内的时间差，因为时间差不太可能超过一年（域名可能会在一年后过期并更改所有权）。 x 轴上的正值表示 Gossip 可以比黑名单更早地检测到恶意域。我们看到 Gossip 比现有的公共黑名单更早检测到 79% 的恶意域。图 3 显示了 30 天（一个月）内的截断分布。大约 40% 的域可以在它们出现在公共黑名单之前 10 天被检测到。因此，Gossip 有更好的机会及早发现攻击，并为用户提供更多时间来缓解威胁。

4.3 公共邮件列表的实验

为了评估 Gossip 的有效性，我们结合公共邮件列表构建了一个综合数据集。公共邮件列表通常不包含恶意行为

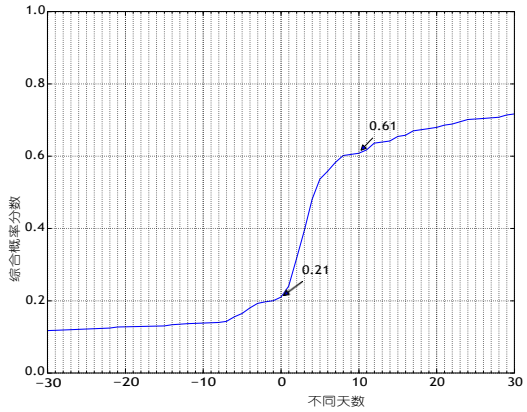

图 3：Gossip 和公共黑名单之间的检测时间差异分布（截断，一个月周期）。

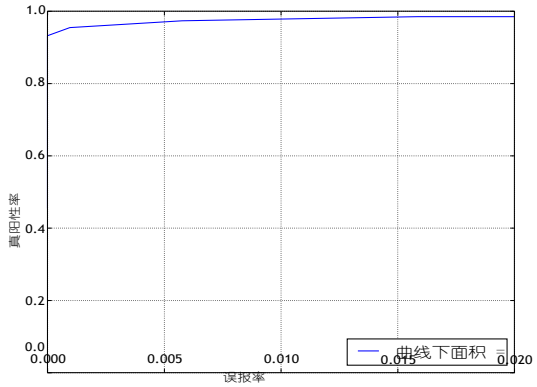

图 4：Gossip 的 ROC 曲线（10 折交叉验证）。

mains，因为他们的主题是关于一般计算机相关问题（不关注安全或攻击）。我们将公共邮件列表中讨论的域提取为良性集，并删除出现在黑名单中的任何域。我们使用安全邮件列表中列入黑名单的域作为恶意集。最后，我们有 693 个恶意域和 3,898 个良性域。

大多数特征可以像以前一样提取。一种特殊情况是参与者特征。安全邮件列表和公共邮件列表不太可能有重叠的用户（具有相同的电子邮件地址）。因此，我们将参与者特征扩展为电子邮件地址中的二级域部分（而不是使用整个电子邮件地址作为特征）。我们使用 10 折交叉验证来评估检测精度。这些域被随机分成 10 个大小相等的部分。在每次迭代中，将其中一部分作为测试数据，其他部分作为训练数据。最终结果是所有迭代的平均值。如前所述，我们使用随机森林作为分类器。

算法 1 计算黑名单和八卦检测之间的时间差。

要求：

d： 目标域；
BT ： 多黑名单提要（MD）中出现时间段的集合；
vt： 首次出现在 VirusTotal 中的时间； mt：在邮件列表中首次出现的时间； dt：域的不同时间；

```
1: 蓝牙 ← 在MD中搜索所有出现时间的d
2: 如果 BT = 空则
3:   如果存在 (ibt ∈ 英国电信）& 第一次 ≤ 公吨 ibt.last.time 然后
4:     德语 分钟 (ibt.first.time
     公吨）5:  ∈ 否则如果存在
   (ibt 英国电信）& 第一次 > 然后山
6:     德语 分钟 (ibt.first.time
     公吨）7:  ∈ 否则如果存在
   (ibt 英国电信）& 第一次 ≤ 然后山
8:     德语 分钟 (ibt.first.time
     mt）9: ∈ 否则如果 d
     然后
10:  德语 vt 公吨
11: 返回dt
```

图 4 显示了 10 折交叉验证的 ROC 曲线。八卦达到了很高的准确性，即 94% 的检测率和零误报。图中曲线下面积（AUC）约为0.99。我们的结果表明，八卦可以跨邮件列表部署，并根据在线讨论有效区分恶意域。

5. 分析与测量

根据上一节的结果，我们执行进一步分析以调查误报，评估最重要的特征，并描述我们在邮件列表讨论中发现的恶意域的类型。

5.1 误报分析

在 4.2.1 节中，我们看到即使 Gossip 能够从安全邮件列表中提取许多恶意域，它也会将一定数量的良性域归类为恶意域（即误报）。我们调查了误报，发现其中大部分是安全博客或公司网站。原因是在讨论恶意活动时，安全专家经常将外部博客或技术网站作为参考以查找更多详细信息。这些良性域在我们的系统中与所讨论的恶意域具有相似的特征，因此有一定的概率被错误分类。如果我们能够将这些与安全相关的域列入白名单，则可以进一步提高 Gossip 的性能。启发式是恶意网站不会显示与安全相关的词，例如“恶意软件”或“威胁”，以提醒访问者。另一方面， 与安全相关的网站在提到恶意活动时经常包含这样的词。因此，我们开发了一个词袋模型来识别安全相关网站上的关键字，然后抓取我们数据集中的域以将良性域列入白名单。

表 3：选择安全网站作为提取安全相关词的种子。

不。	网站网址
1	http://lastline.com
2	https://usa.kaspersky.com
3	https://www.fireeye.com
4	https://www.damballa.com
5	http://home.f-secure.com
6	https://www.lancope.com
7	https://www.alienvault.com
8	http://www.avg.com
9	http://www.trendmicro.com
10	https://www.paloaltonetworks.com

我们选择了 10 个知名的安全网站（如表 3 所列）作为提取安全相关词的种子。算法 2 解释了使用自然语言处理技术生成安全关键词（第 1-8 行）并按词频（TF）和逆文档频率（IDF）算法（第 9-14 行）对这些关键词进行排名的详细提取步骤。为了找到最能代表安全主题的关键字，我们计算了每个词的 TF-ID 值，其中 IDF 值是从 Microsoft 的 N-gram 服务 [47] 计算得出的。我们根据TF-IDF值对词进行排序，选取前50个词生成安全关键词语料库。

我们实验中与安全相关的前 50 个关键字是：“security”、“threat”、“amp”、“avg”、“kaspersky”、“lastline”、“business”、“protection”、“cyber”、“2015”，“端点”，“网络”，“高级”，“情报”，“学习”，“alienvault”，“恶意软件”，“产品”，“趋势”，“企业”，“获取”，“威胁”，“antivirus”，“mobile”，“partner”，“fireeye”，“internet”，“free”，“support”，“data”，“micro”，“damballa”，“detection”，“management”，“home”，“小”，“攻击”，“兰科普”，“检测”，“联系”，“新”，“事件”，“防御”，“在线”，“联合”，“交付”，“突破”，“零日”、“病毒”、“服务”。

接下来，我们抓取数据集中的域并从他们的网站中提取单词。如果一个网站的安全关键词语料库中至少有N个词，我们认为对应的域是安全相关域（适用于白名单）。我们在实验中将 N 设置为 5。

我们重新审视了 Goss 输出的人工检查过程。表 4 显示了与表 2 相比的白名单结果。括号中的数字表示白名单可以过滤的域数。特别是，对于排名较低的域，如 1,460-1,559 组中的域，相当一部分被错误分类的良性域可以列入白名单。通过白名单，Gossip 可以达到更高的检测准确率。

5.2 特征重要性

通过分类结果，我们确定了 Gossip 中最有效的特征是什么。这种洞察力可以帮助未来的功能开发，并且可能会刺激邮件列表或 Web 论坛提供结构化格式或接口来提取有用的功能。我们通过使用平均减少杂质（MDI）方法 [38] 来评估特征重要性。表 5 显示了我们系统中 10 个最有效的功能。在前五个特征中，有四个属于 URL 特征和 Thread 特征，需要较小的开销来提取。

算法2 利用TF-IDF算法提取安全网站关键词

要求：

D：一组测试域；
T W ：来自所有网站的词集； T FIDF ：所有单词的 td idf 值集； RV ：前 50 个安全关键字的集合；

1：对于每个 d ∈ D做
2： W ⇐ 从 html(d) 中提取文本；
3： W ⇐ 单词标记化 (W)；
4： W ⇐ 删除停用词 (W)； 5：
对于每个 w ∈ 做
6： 如果 len(w) [2, 20] 那么
7： w ⇐ 词干 (w.lower ())；
8： TW.append(w);
9: 对于每个 tw ∈ 做
10： tf ⇐ T W.counter(tw)/len(T W)
11： 以色列国防军 _ _
微软 Ngram idf (tw) 12： T FIDF.append ((tf idf, * tw), tw)
13： T FIDF ⇐ 按值排序 (T FIDF)； 14：房车 T FIDF.get most 50 words; 15: 返回房车;

表 4：与表 2 相比的白名单结果。括号中的数字表示白名单域的数量。

秩 地区	列入黑名单	非黑名单		
		恶意	良性	不确定
1-100	60	36	4 (**-2**)	0
101-200	62	29	8 (**-3**)	1 (**-1**)
...	...	...	...	...
1,460-1,559	20	22 (**-4**)	55 (**-31**)	3 (**-1**)
...	...	...	...	...

为了显示有关功能排名的完整图片，我们列出了接下来的 40 个功能。上下文特征中的关键字用引号引起来。由于参与者特征包含个人信息，我们只显示特征索引。排名特征列表是：“fw”，“folk”，“server”，“more”，“associate”，已知恶意软件检测服务的数量，“intel”，“hash”，“hxxp”，域数量，“public”，参与者特征 #456，“reject”，“refer”，“lot”，“korean”，“marco”，“jeroen”，IP 地址数，“down”，“origin”，“regard”，特征 #132，“事件”，“入侵”，“iframe”，参与者特征 #387，“poc”，参与者特征 #787，“uri”，“firm”，“advance”，“investigate”，参与者特征 #1287，“against”，“attack”，“jose”，participant feature #21，“myself”。

5.3 邮件列表中的域特征

我们提供邮件列表中讨论的恶意域的高级统计数据，以了解哪些恶意活动引起最多的讨论，以及它们来自何处。我们手动检查了采样的恶意域（参见第 4.2.1 节），并将它们分为不同的类别。在表 6 中，第一列显示了恶意活动的类别，第二列是该类别中的域占所有恶意域的百分比。我们看到路过式下载攻击、C&C 服务器和恶意软件下载服务器所占的比例最大。大概是因为这些攻击更加隐蔽

表 5：八卦中最重要的 10 个特征。

不。	特征	特征组
1	Alexa排名	网址功能
2	"n' i"	上下文特征
3	回复数	螺纹特征
4	附件数量	螺纹特征
5	**“**某物**”**	上下文特征
6	以 .exe 结尾的网址	网址功能
7	**“**相信**”**	上下文特征
8	"稳"	上下文特征
9	组织数量	螺纹特征
10	密码摘要的数量	螺纹特征

表 6：邮件列表中讨论的恶意活动类别分布。

类别	域百分比
路过式下载	34.57%
C&C	22.84%
恶意软件下载服务器	17.28%
网络钓鱼	9.88%
网络后门	6.17%
垃圾邮件	5.56%
DDoS僵尸网络	3.70%

（即不易检测）并造成直接损害，安全专家倾向于在邮件列表中更频繁地讨论它们。

接下来，我们调查 DNS 基础设施以及哪些父区域托管最恶意的域。请注意，在我们的分析中，一些父区域是顶级域，一些是二级域（如第 2.2 节所述）。

表 7 显示了父区域的分布。 “.com”区域仍然是网络犯罪分子用于非法活动的主要目标。在国家代码顶级域中，**“**.ru”区域是被滥用最多的区域。

6. 局限性

与其他采用机器学习技术的恶意软件检测系统一样，我们的方法在应用于邮件列表时存在一些局限性。在本节中，我们将讨论这些限制，并提出 Gossip 如何在实际部署中更有效地工作。

邮件列表中的大多数线程都包含对帖子的许多讨论，但也有一些线程回复很少或字数有限。如果在回复或字数很少的帖子中提到任何恶意域，则很难正确分类该域。这种情况导致更高的假阴性率。这种局限性已在之前的研究中得到研究 [36]。如果训练数据量足够大，或者如果可以主动抓取域以获取更多信息，则风格特征（例如推理连词）可能有助于提高预测的准确性。

另一个问题是某些线程可能包含大量域并且内容中包含的单词很少，例如，当人们讨论许多解析为同一 IP 地址的域时。如果训练数据集包含许多此类实体，则检测系统的效率会降低。但是，避免此问题的解决方案是删除此类邮件线程

表 7：邮件列表中讨论的恶意域的父区域分布。

区	域覆盖	区	域覆盖
。.com	47.66%	com.tw	0.87%
网	7.31%	我们	0.87%
组织	6.32%	在	0.87%
茹	5.78%	通道	0.76%
信息	2.84%	教育	0.65%
de	2.18%	英国公司	0.65%
康	1.96%	辅佐	0.65%
欧洲联盟	1.64%	在	0.65%
pl	1.64%	co.kr	0.65%
澳洲网	1.42%	传统知识	0.55%
商业	1.31%	是	0.55%
荷兰	1.09%	其他的	11.13%

在训练数据集中，或者只是从相同的线程中提取一定数量的域。还存在其他缓解方法，例如，先前的研究从 URL 中提取了更多特征（长度、数字等）[48]，Xiang 等人。从实体的网页和 WHOIS 信息中提出了更多功能 [50]。但是这样的方式需要更多的网络请求和及时的响应来抓取网页。

我们使用自然语言处理技术来提取上下文特征，如第 3 节所示，但由于国家和文化的多样性，人们可能会使用不同的单词和句子，并且可能会在线程中出现不同的书写错误和语法错误。萨迪亚等人。已经使用文体计量方法通过分析他们的写作风格来识别匿名作者 [7]。为了解决这些问题，人们可以采用更复杂的语言学方法来提取上下文特征。

7. 相关工作

恶意软件检测是安全领域的一个普遍问题，引起了广泛的研究兴趣。不同的研究人员使用各种方法来研究这个主题，例如动态分析、URL 中的词法模式以及地下市场的分析。在下文中，我们重点介绍了最相关的先前工作。

恶意软件检测。有几种特征可以表明恶意网站和良性网站之间的区别：网页中的可疑代码、沙箱中的浏览器行为或域元数据。这些用于检测恶意域的方案通常用于机器学习技术。这些特征可以基于网站内容的静态特征[12、16、27]，也可以从它们的动态行为中提取[14、15、28]。域注册信息和 DNS 流量也提供了识别恶意域的重要模式 [9、10、11、23、24、25、51]，但需要使用外部资源或在特定的网络有利位置提取此类特征。在特征提取之后，可以使用各种机器学习分类器对这些实体进行分类。

列入黑名单。有许多研究着眼于黑名单：K¨hrer 等人。提出了一个黑名单解析系统来跟踪 49 个不同的黑名单，并提供了对源自这些公众的恶意活动的分析黑名单 [29]。 K¨hrer 发现 15 个公共黑名单包含不到 20% 的大多数流行恶意软件家族的恶意域 [30]。马等。描述了一种基于机器学习方法使用词法和基于主机的特征对恶意和良性 URL 进行分类的方法 [32]。 Almuhimedi 等人。研究了人类对谷歌浏览器 [8] 中恶意软件警告的反应行为。此类研究表明，黑名单已被广泛用于阻止恶意域和 IP 地址，但也概述了黑名单的几个不足之处。廖等。提出了仅基于上下文术语的全自动妥协指标（IOC）提取的解决方案[31]，但邮件列表讨论有不同的情况，其中技术描述的单词较少，同一主题有多个讨论。因此，他们的解决方案无法奏效。

自然语言处理技术。 NLP 技术最近在安全领域得到广泛应用，尤其是在检测网络钓鱼和恶意域方面：Darling 等人。开发了一个基于 URL 词汇特征的分类系统 [18]。其他研究人员还研究了使用词汇和分词技术来提取各种特征（例如域名长度、某些字符的出现频率）的恶意域检测 [26、34、46]。 N-Gram 方法也被用于检测未知网络攻击 [42]。

与之前的研究相比，我们的方法使用上下文词和机器学习技术来检测邮件列表中的恶意域。这项研究可以分析传统检测解决方案遗漏的可疑域，因为许多恶意域只出现几次并且持续时间很短。我们还根据域和安全专家之间的历史关系提出参与者特征。最后，我们使用自然语言处理技术来提取基于上下文词的特征，这些特征以前没有在该领域中使用过。

8. 结论

在本文中，我们设计了 Gossip，一种新颖的轻量级检测系统，用于从邮件列表中提取恶意域。我们确定了四组特征：URL 特征、线程特征、参与者特征和上下文特征。我们将这些特征合并到分类器中以开发检测系统。为了为每个领域派生出有意义的关键词，我们引入了一种改进的连续词袋模型，该模型可以提取所讨论领域周围的代表性词。为了过滤邮件列表中的安全公司网站或技术博客，我们提出了一种从网页中抓取关键字并减少误报的方法。我们用真实数据评估了 Gossip。我们的结果表明，八卦通过私人邮件列表实现了高精度，并捕获了现有公共黑名单遗漏的数百个恶意域。此外，我们的系统比现有的公共黑名单更早检测到 79% 的恶意域。 通过在公共邮件列表上的实验，Gossip 获得了 94% 的检测率和零误报，这也证明了我们模型的可靠性和准确性。

参考文献

[1] Malc0de 数据库。http://malc0de.com/database.

[2] 恶意软件域列表。http://www.malwaredomainlist.com.

[3] Mozilla 公共足够列表。http://publicsuffix.org.

[4] 网络钓鱼坦克。 https://www.phishtank.com。

[5] 垃圾邮件之家项目。https://www.spamhaus.org。

[6] 瑞士安全博客。https://www.abuse.ch.

[7] S. Afroz、A. C. Islam、A. Stolerman、R. Greenstadt 和 D. McCoy。 Doppelg¨nger finder：将文体测量法带到地下。在 IEEE 安全和隐私研讨会上，2014 年。

[8] H. Almuhimedi、A. P. Felt、R. W. Reeder 和 S.康索沃。你的声誉先于你：历史、声誉和 Chrome 恶意软件警告。在可用隐私和安全（SOUPS）研讨会上，2014 年。

[9] M. Antonakakis、R. Perdisci、D. Dagon、W. Lee 和 N.费姆斯特。为 DNS 构建动态信誉系统。在第 19 届 USENIX 安全研讨会论文集中，2010 年。

[10] M. Antonakakis、R. Perdisci、W. Lee、N. Vasiloglou II 和 D. Dagon。在上层 DNS 层次结构中检测恶意软件域。在第 20 届 USENIX 安全研讨会论文集中，2011 年。

[11] L. Bilge、E. Kirda、C. Kruegel 和 M. Balduzzi。暴露：使用被动 DNS 分析查找恶意域。在网络和分布式系统安全研讨会（NDSS）的会议记录中，2011 年。

[12] D. Canali、M. Cova、G. Vigna 和 C. Kruegel。Prophiler：用于大规模检测恶意网页的快速过滤器。在国际万维网会议（WWW）的会议记录中，2011 年。

[13] R. Caruana 和 A. Niculescu-Mizil。监督学习算法的实证比较。在第 23 届国际机器学习会议论文集中，2006 年。

[14] 厘米。陈，J.-J。黄和 Y.-H。欧。基于域统计信息检测 Web 攻击。在情报和安全信息学中，第 97-106 页。施普林格，2013 年。

[15] M. Cova、C. Kruegel 和 G. Vigna。检测和分析驱动下载攻击和恶意 javascript 代码。在万维网会议（WWW）的会议记录中，2010 年。

[16] C. Curtsinger、B. Livshits、B. Zorn 和 C. Seifert。 Zozzle：低开销的主要是静态 Javascript 恶意软件检测。在第 20 届 USENIX 安全研讨会论文集中，2011 年。

[17] *M.卡茨。牛津简明英语指南。牛津大学出版社，2013 年。*

[18] M. Darling、G. Heileman、G. Gressel、A. Ashok 和 *P. Poornachandran。一种用于对恶意 URL 进行分类的词法方法。在 IEEE 高性能计算与模拟国际会议（HPCS），第 195-202 页，2015 年。*

[19] G. 团体。 10,000 个最常用的英语单词。https://github.com/first20hours/google-10000-english。

[20] G. 团体。病毒总数。https://www.virustotal.com.

[21] R. Gunning 等人。如何消除写作中的迷雾。1964.

[22] N. Habash、O. Rambow 和 R. Roth。马达+令牌：用于阿拉伯语标记化、变音符号化、词法消歧、词性标记、词干提取和词形还原的工具包。在第二届阿拉伯语语言资源和工具（MEDAR）国际会议论文集中，2009 年。

[23] S. Hao、N. Feamster 和 R. Pandrangi。监控恶意域的初始 DNS 行为。在 *ACM 互联网测量会议论文集，2011 年。*

[24] S. Hao、A. Kantchelian、B. Miller、V. Paxson 和 N.费姆斯特。掠夺者：在注册时主动识别和消除域名滥用。在 ACM 计算机和通信安全会议上，2016 年。

[25] S.郝，M.托马斯，V.帕克森，N.费姆斯特，C. Kreibich、C. Grier 和 S. Hollenbeck。了解垃圾邮件发送者的域注册行为。在 ACM 互联网测量会议论文集中，2013 年。

[26] *Y. He、Z. Zhong、S. Krasser 和 Y. Tang。为恶意域注册挖掘 DNS。第六届国际协作计算会议论文集：网络、应用和工作共享，2010 年。*

[27] L.因韦尔尼齐，P.M.孔帕雷蒂，S.本韦努蒂，C. Kruegel、M. Cova 和 G. Vigna。 EvilSeed：一种查找恶意网页的指导方法。在 IEEE 安全和隐私研讨会上，2012 年。

[28] A. Kapravelos、M. Cova、C. Kruegel 和 G. Vigna。逃离猴子岛：逃避高交互蜜客户端。在第 8 届入侵检测和恶意软件与漏洞评估（DIMVA）会议记录中。 2011.

[29] *M.K.赫勒和T.霍尔茨。恶意软件黑名单的实证分析。信息与通信实践， 35（1）：11-16，2012.*

[30] M. K¨hrer、C. Rossow 和 T. Holz。将其涂黑：评估恶意软件黑名单的有效性。在入侵检测最新进展研讨会上。 2014.

[31] X. Liao、K. Yuan、X. Wang、Z. Li、L. Xing 和 R.贝亚。参加国际奥委会比赛：实现开源网络威胁情报的自动发现和分析。在 ACM 计算机和通信安全会议上，2016 年。

[32] J. Ma、L. K. Saul、S. Savage 和 G. M. Voelker。超越黑名单：学习从可疑 URL 检测恶意网站。在第 15 届 ACM SIGKDD 知识发现和数据挖掘会议论文集中，2009 年。

[33] 迈克菲。https://www.siteadvisor.com.

[34] D. K. McGrath 和 M. Gupta。网络钓鱼的背后：对网络钓鱼者作案手法的检查。在 USENIX 大规模漏洞利用和紧急威胁（LEET）研讨会论文集中，2008 年。

[35] T. 摩尔和 R. 克莱顿。评估人群在评估网络钓鱼网站时的智慧。在金融密码学会议论文集和 *数据安全。 2008.*

[36] S. Mukherjee、G. Weikum 和 C. Danescu-Niculescu-Mizil。吸毒者：健康社区用户陈述的可信度。

在第 20 届 ACM SIGKDD 知识发现和数据挖掘会议论文集中，2014 年。

[37] A. Narayanan、H. Paskov、N. Z. Gong、J. Bethencourt、
E. Stefanov、E. C. R. Shin 和 D. Song。论互联网规模作者识别的可行性。在 IEEE 安全和隐私研讨会上，2012 年。

[38] L. Olshen、C. J. Stone 等人。分类和回归树。沃兹沃斯国际集团，93(99):101，1984。

[39] A.皮西利迪斯，C.卡尼奇，G.M.沃尔克，
K. Levchenko 和 S. Savage。品尝者的选择：垃圾邮件提要的比较分析。在 ACM 互联网测量会议论文集中，2012 年。

[40] M.F.波特。 Snowball：一种用于词干提取算法的语言，2001 年。

[41] L. 理查森。漂亮的汤文档。 2007.

[42] *K. Rieck 和 P. Laskov。使用语言模型检测未知网络攻击。在入侵和恶意软件检测以及漏洞评估国际会议上。 2006.*

[43] S. Sheng, B. Wardman, G. Warner, L. F. Cranor,
J. Hong 和 C. Zhang。网络钓鱼黑名单的实证分析。在第六届电子邮件和反垃圾邮件 (CEAS) 会议记录中，2009 年。

[44] B. Steven、E. Klein 和 E. Loper。使用 Python 进行自然语言处理。奥莱利媒体，2009 年。

[45] P. Vadrevu、B. Rahbarinia、R. Perdisci、K. Li 和
M. Antonakakis。测量和检测实时网络流量中的恶意软件下载。在欧洲计算机安全研究研讨会论文集中，2013 年。

[46] K. Wang、C. Thrasher 和 B.-J。 P.许。 Web scale NLP：关于 URL 断词的案例研究。在第 20 届万维网国际会议论文集中，2011 年。

[47] K. Wang、C. Thrasher、E. Viegas、X. Li 和 B.-j。 P.许。 Microsoft Web N-gram 语料库和应用程序概述。在 NAACL HLT 2010 演示会议记录中，第 45-48 页。计算语言学协会，2010 年。

[48] W. Wang 和 K. E. 雪莉。绝命毒师：使用分词检测恶意域。在 IEEE Web 2.0 安全和隐私研讨会中。2015.

[49] P.威利特。波特词干提取算法：过去和现在。计划，40(3):219－223，2006。

[50] G. Xiang、J. Hong、C. P. Rose 和 L. Cranor。Cantina+：一个功能丰富的机器学习框架，用于检测网络钓鱼网站。 ACM 信息和系统安全交易 (TISSEC)，14(2):21，2011 年。

[51] W. Zhang、W. Wang、X. Zhang 和 H. Shi。 DNS WHOIS信息隐私保护研究在计算机科学及其应用中，第 71-76 页。施普林格，2015 年。