高精度检测商务邮件入侵

Asaf Cidon、Barracuda Networks 和哥伦比亚大学； Lior Gavish、Itay Bleier、Nadia Korshun、Marco Schweighauser 和 Alexey Tsitkin，Barracuda Networks

https://www.usenix.org/conference/usenixsecurity19/presentation/cidon

这篇论文包含在第 28 届 USENIX 安全研讨会论文集中。

2019 年 8 月 14-16 日 • 美国加利福尼亚州圣克拉拉

978-1-939133-06-9

开放获取第 28 届 USENIX 安全研讨会论文集
由 USENIX 赞助。

高精度检测商务邮件入侵

Asaf Cidon[1,2]和 Lior Gavish、Itay Bleier、Nadia Korshun、Marco Schweighauser 和 Alexey Tsitkin[1]1Barracuda Networks, [2]哥伦比亚大学

摘要

商业电子邮件泄露（BEC）和员工冒充已成为成本最高的网络安全威胁之一，造成超过 120 亿美元的报告损失。冒充电子邮件有多种形式：例如，有些要求电汇到攻击者的帐户，而另一些则引导收件人点击链接，这会危及他们的凭据。电子邮件安全系统无法有效检测这些攻击，因为这些攻击不包含明显的恶意负载，并且针对收件人进行了个性化设置。

我们介绍了 BEC-Guard，这是一种在 Barracuda Networks 使用的检测器，它使用监督学习实时防止商业电子邮件泄露攻击。 BEC-Guard 自 2017 年 7 月开始投入生产，是 Barracuda Sentinel 电子邮件安全产品的一部分。 BEC-Guard 依靠有关可通过云电子邮件提供商 API 访问的历史电子邮件模式的统计数据来检测攻击。设计 BEC-Guard 时的两个主要挑战是需要标记数百万封电子邮件以训练其分类器，以及在员工冒充电子邮件的发生非常罕见时正确训练分类器，这可能会使分类产生偏差。我们的主要见解是将分类问题分为两部分，一部分分析电子邮件的标题，第二部分应用自然语言处理来检测与 BEC 相关的短语或电子邮件正文中的可疑链接。 BEC-Guard 利用云电子邮件提供商的公共 API 自动学习每个组织的历史通信模式，并实时隔离电子邮件。我们在包含 4,000 多次攻击的商业数据集上评估了 BEC-Guard，结果表明它达到了 98.2% 的准确率和低于五百万分之一电子邮件的误报率。

1 介绍

近年来，被 FBI 称为“商业电子邮件妥协”(BEC) 的电子邮件员工冒充已成为主要的安全威胁。据 FBI 称，美国组织在 2018 年累计损失 27 亿美元自 2013 年以来 120 亿美元 [13].众多知名企业已成为此类攻击的牺牲品，包括 Facebook、谷歌 [41]，以及泛在[44].研究表明，与勒索软件等其他常见网络攻击相比，BEC 造成的直接经济损失要高得多 [11, 13]. BEC 攻击还诱骗了关键政府基础设施的运营商 [39].就连消费者也成了员工冒充的对象。例如，攻击者冒充房地产公司的员工，诱骗购房者将首付款电汇到错误的银行账户 [1, 7, 17].

BEC 有多种形式：一些电子邮件要求收件人将钱电汇到攻击者的帐户，其他电子邮件要求提供包含社会安全号码的 W-2 表格，还有一些会引导收件人点击钓鱼链接，以窃取他们的凭据。共同的主题是冒充目标的经理或同事 [12].在这项工作中，我们重点关注攻击者在组织外部并试图冒充员工的攻击。在 §6 我们讨论其他场景，例如攻击者使用受损的内部电子邮件帐户冒充员工 [18,19].

大多数电子邮件安全系统无法有效检测 BEC。在分析传入的电子邮件时，电子邮件安全系统广泛地寻找两种类型的属性：恶意的和容量大的。恶意属性的示例包括包含恶意软件的附件、指向受感染网站的链接或从信誉不佳的域发送的电子邮件。有各种众所周知的技术来检测恶意属性，包括沙盒 [49]，以及域名声誉[2, 48].当将相同的电子邮件格式发送给数百个或更多收件人时，会检测到体积属性。示例包括相同的文本或发件人电子邮件（例如，垃圾邮件）和相同的 URL（例如，大规模网络钓鱼活动）。然而，员工冒充电子邮件不包含恶意或容量属性：它们通常不包含恶意软件，不是从众所周知的恶意 IP 发送的，通常不包含链接，并且发送给少数收件人（带有明确的逃避体积过滤器的意图）。当员工冒充攻击确实包含链接时，通常是

指向合法网站上已被破坏的虚假注册页面的链接，该链接未出现在任何 IP 黑名单中。此外，攻击文本是为接收者量身定制的，通常不会被基于容量的过滤器捕获。

我们的设计目标是以低误报率（百万分之一的电子邮件）和高精度（95%）实时检测和隔离 BEC 攻击。我们观察到流行的云电子邮件系统（如 Office 365 和 Gmail）提供的 API 使帐户管理员能够允许外部应用程序访问历史电子邮件。因此，我们设计了一个系统，依靠通过这些 API 提供的历史电子邮件来检测 BEC。

之前检测假冒的工作是在非常小的数据集上进行的[10, 14, 20, 45])，或专注于阻止 BEC 攻击的一个子集（域欺骗［14］或带有链接的电子邮件［20］).此外，大多数先前的工作都存在精度非常低的问题（500 次警报中只有 1 次是攻击［20］）或非常高的误报率［10, 45]），这使得之前的工作不适合实时检测 BEC。

设计一个能够以低误报率检测 BEC 的系统的主要挑战是 BEC 电子邮件在所有电子邮件中所占的百分比非常少。事实上，在我们的数据集中，不到 50,000 封电子邮件中有一封是 BEC 攻击。因此，为了实现低误报率，我们设计了一个使用监督学习的系统，该系统依赖于大量的 BEC 电子邮件训练集。然而，引导监督学习系统存在两个实际挑战。首先，很难标记包含数百万封电子邮件的足够大的训练数据集。其次，在不平衡数据集上训练分类器具有挑战性，其中训练数据集包含的正样本（即 BEC 攻击）比负样本（即无辜电子邮件）少近五个数量级。

在本文中，我们介绍了我们最初是如何训练 BECGuard 的，BECGuard 是一个安全系统，可以使用历史电子邮件实时自动检测和隔离 BEC 攻击。 BECGuard 是商业产品 Barracuda Sentinel 的一部分，Barracuda Networks 的数千家企业客户使用该产品来防止 BEC、帐户接管、鱼叉式网络钓鱼和其他有针对性的攻击。 BEC-Guard 不需要分析师审查检测到的电子邮件，而是依赖于离线和不频繁的分类器重新训练。BEC-Guard 的关键见解是将训练和分类分为两部分：header 和 body。

模拟分类器不是直接对 BEC 攻击进行分类，而是通过检查电子邮件标题来确定攻击者是否在冒充公司员工，从而检测模拟尝试。它利用的功能包括员工通常使用哪些电子邮件地址、他们的名字有多受欢迎以及发件人域的特征等信息。内容分类器仅在被归类为模拟尝试的电子邮件上运行，并检查电子邮件正文中的 BEC。对于不包含链接的电子邮件，我们使用 k-最近邻[43]（KNN）分类器，使用词频对词进行加权-逆文档频率[28, 42]（TFIDF）。对于带有链接的电子邮件，我们训练了一个随机森林分类器，该分类器依赖于受欢迎程度以及链接在文本中的位置。可以使用客户反馈频繁地重新训练这两个内容分类器。

为了创建初始分类器，我们分别标记和训练每种类型的分类器：模拟分类器的标签是使用我们在训练数据集上运行的脚本生成的，而内容分类器是在手动标记的训练数据集上训练的。由于我们仅对检测为假冒尝试的电子邮件运行内容分类，因此我们需要手动标记训练数据集的一个小得多的子集。此外，为了确保模拟分类器在不平衡数据集上成功训练，我们使用高斯混合模型（一种无监督聚类算法）为合法电子邮件开发了欠采样技术。分类器通常每隔几周重新训练一次。可用于初始培训的数据集包括来自 1500 位客户一年的历史电子邮件，总数据集包含 200 万个邮箱和 25 亿封电子邮件。自训练初始分类器以来， 我们的数据集已经扩展到包括数千万个邮箱。

BEC-Guard 使用基于云的电子邮件系统（例如 Office 365 和 Gmail）的 API，既可以在数小时内自动了解每个组织的历史通信模式，又可以实时隔离电子邮件。 BEC-Guard 订阅 API 调用，每当新电子邮件进入组织的邮箱时，它会自动提醒 BEC-Guard。一旦收到 API 调用的通知，BEC-Guard 就会对电子邮件进行 BEC 分类。如果电子邮件被确定为 BEC，BEC-Guard 使用 API 将电子邮件从收件箱文件夹移动到最终用户帐户上的专用隔离文件夹。

为了评估我们方法的有效性，我们在从数百个组织获取的电子邮件数据集上测量了 BEC-Guard 的性能。在这个标记的数据集中，BEC-Guard 达到了 98.2% 的精度，误报率仅为 530 万分之一。总而言之，我们做出以下贡献：

- 首个高精度、低误报率的实时BEC防范系统。
- BEC-Guard 的新颖设计依赖于云电子邮件提供商 API 来了解每个组织的历史通信模式，并实时检测攻击。为了应对标记数百万封电子邮件，我们将检测问题分成两组顺序运行的分类器。
- 我们对电子邮件的标题和文本使用不同类型的分类器。标题使用随机森林进行分类，而文本分类主要依赖于不依赖于任何硬编码特征的 KNN 模型，并且可以轻松地重新训练。
- 为了在不平衡数据集上训练模拟分类器，我们使用聚类算法对合法电子邮件进行抽样。

BEC目标	关联？	百分比
电汇	不	46.9%
点击链接	是的	40.1%
建立良好关系	不	12.2%
窃取 PII	不	0.8%

表 1： BEC 攻击的目标占 3,000 次随机选择的攻击的百分比。59.9% 的攻击不涉及网络钓鱼链接。

角色	接受者 %	模拟百分比
首席执行官	2.2%	42.9%
首席财务官	16.9%	2.2%
C级	10.2%	4.5%
财务/人力资源	16.9%	2.2%
其他	53.7%	48.1%

表 2： 从 50 家随机公司中选择的 BEC 攻击样本中的接收者和假冒员工的角色。 Clevel 包括除 CEO 和 CFO 之外的所有高管，而财务/人力资源不包括高管。

2 背景

商业电子邮件泄露，也称为员工冒充、CEO 欺诈和捕鲸，[1] 是一类电子邮件攻击，攻击者冒充公司员工（例如，首席执行官、人力资源或财务经理），并制作一封发送给特定员工的个性化电子邮件。这封电子邮件的目的通常是诱骗目标汇款、发送敏感信息（例如 HR 或医疗记录），或引导员工点击网络钓鱼链接以窃取他们的凭据或将恶意软件下载到他们的端点。

BEC 已成为近年来最具破坏性的电子邮件攻击之一，与垃圾邮件和勒索软件等其他类型的攻击持平或超过。由于 BEC 攻击的严重性，FBI 开始根据向 FBI 报告欺诈性电汇的美国组织编制年度报告。根据 FBI 的数据，2013 年至 2018 年间，损失了 120 亿美元 [13]. 为了正确看待这一点，谷歌的一项研究估计，2016 年勒索软件支付的总额仅为 2500 万美元 [11].

在本节中，我们回顾了 BEC 的常见示例，并提供了关于如何利用它们的独特特征进行监督学习分类的直觉。

2.1 统计数据

为了更好地理解 BEC 攻击的目标和方法，我们对数据集中 3,000 次随机选择的 BEC 攻击进行了统计（有关我们数据集的更多信息，请参阅 § 4.2）. 桌子1 总结了攻击的目标。结果表明，样本攻击中最常见的 BEC 是试图欺骗收件人向攻击者拥有的银行账户进行电汇，而约 0.8% 的攻击要求收件人向攻击者发送电汇个人身份信息（PII），通常采用包含社会安全号码的 W-2 表格形式。大约 40% 的攻击要求接收者单击链接。12% 的攻击试图通过与接收者开始对话来与目标建立融洽关系（例如，攻击者会询问接收者他们是否有空执行紧急任务）。对于“rapport”邮件，在绝大多数情况下，在回复初始邮件后，攻击者会要求进行电汇。

一个重要的观察是，大约 60% 的 BEC 攻击不涉及链接：攻击只是一封纯文本电子邮件，欺骗收件人进行电汇或发送敏感信息。这些纯文本电子邮件对于现有的电子邮件安全系统以及检测之前的学术工作来说尤其困难[20]，因为它们通常是从合法的电子邮件帐户发送的，针对每个收件人量身定制，并且不包含任何可疑链接。

我们还在我们的数据集中对来自 50 家随机公司的攻击进行了抽样，并对攻击接收者和冒充发送者的角色进行了分类。桌子2 呈现结果。根据结果，用于描述 BEC 的“CEO 欺诈”一词确实是有道理的：大约 43% 的冒充发件人是 CEO 或创始人。攻击的目标在不同角色之间的分布更加平均。然而，即使是冒充的发件人，大多数（约 57%）也不是 CEO。几乎一半的冒充角色和超过一半的目标都不是“敏感”职位，例如高管、财务或人力资源。因此，仅仅保护敏感部门的员工不足以防范 BEC。

2.2 BEC的常见类型

为了引导讨论，我们描述了数据集中 BEC 攻击的三个最常见示例：电汇、融洽关系和假冒网络钓鱼。在 §6 我们将讨论本文未涵盖的其他攻击。我们提供的所有三个示例都是来自我们数据集中的真实 BEC 攻击，其中名称、公司、电子邮件地址和链接都已匿名化。

示例 1： 电汇示例

```
来自：“简·史密斯”〈jsmith@acrne.com〉致：
“乔·巴恩斯”〈jbarnes@acme.com〉主题：供应
商付款

嘿，乔，

你在附近吗？我需要尽快向供应商发送电汇。

简
```

在示例中1，攻击者要求执行电汇。其他类似请求包括要求提供 W-2 表格、医疗信息或密码。在示例中，攻击者伪造了一名员工的姓名，但使用了一个电子邮件地址

[1] 我们在整篇论文中将这种攻击称为 BEC。

示例 2：融洽示例

来自：“简·史密斯” <jsmith@acme.com>
回复：“简·史密斯” <ceo.executive@outlook.com> 致：
“乔·巴恩斯”<jbarnes@acme.com>
主题：在办公桌？

乔，有急事吗？

示例 3：带有钓鱼链接的仿冒名称

来自：“简·史密斯” <greyowl1234@comcast.net> 致：
“乔·巴恩斯”<jbarnes@acme.com>
主题：发票到期编号 381202214

我今天试图通过电话与您联系，但打不通。请在下方告知
我发票的状态。

发票到期编号 381202214：
[http://firetruck4u.net/past-due-invoice/]

不属于组织的域。一些攻击者甚至使用看起来与目标组织域相似的域（例如，攻击者使用 acrne.com 而不是 acme.com）。由于很多电子邮件客户端不显示发件人电子邮件地址，即使攻击者使用不相关的电子邮件地址，一些收件人也会被欺骗。

例子2 试图营造一种紧迫感。在收件人回复电子邮件后，攻击者通常会要求进行电汇。电子邮件具有员工的发件人地址，而回复地址会将响应转发回攻击者。 DMARC、SPF 和 DKIM 等电子邮件身份验证技术可以帮助阻止欺骗性电子邮件。然而，绝大多数组织不强制执行电子邮件身份验证[25]，因为它可能很难正确实施，并且经常导致合法电子邮件被阻止。[2] 因此，我们的目标是在不依赖 DMARC、SPF 和 DKIM 的情况下检测这些攻击。

例子3 使用欺骗性名称，并试图让收件人点击钓鱼链接。此类网络钓鱼链接通常不会被现有解决方案检测到，因为该链接对收件人来说是唯一的（“零日”）并且不会出现在任何黑名单中。此外，攻击者经常破坏相对信誉良好的网站（例如，小型企业网站）以获取网络钓鱼链接，这些链接通常被电子邮件安全系统归类为高信誉链接。电子邮件中的链接通常会将收件人带到一个网站，在那里他们会被提示登录网络服务（例如，发票应用程序）或下载恶意软件。

3 直觉：利用每次攻击的独特属性

这三个示例都包含独特的特征，这使它们有别于无辜的电子邮件。我们首先去在每个示例的标题中记录独特的属性，然后讨论电子邮件正文的属性以及如何使用它们来构建机器学习分类器的特征。我们还讨论了这些属性的合法极端情况，这些情况可能会欺骗分类器并导致误报。标头属性。在示例中1 和3，攻击者冒充一个人的名字，但使用与公司电子邮件地址不同的电子邮件地址。因此，如果一封电子邮件包含员工的姓名，但使用的电子邮件地址不是该员工的典型电子邮件地址，则发件人很可能是冒名顶替者。

但是，存在员工使用非公司电子邮件的合法用例。首先，员工可能会使用个人电子邮件地址向自己或公司的其他员工发送或转发信息。理想情况下，机器学习分类器应该能够学习属于某个人的所有电子邮件地址，包括公司和个人电子邮件地址。其次，如果外部发件人与内部员工同名，则可能看起来像是冒充。

在示例中2，攻击者欺骗发件人的合法电子邮件地址，但回复电子邮件地址与发件人地址不同，这是不寻常的（我们还将讨论攻击者从发件人的合法地址发送邮件而没有更改 § 中的回复字段6）.然而，这种模式也有合法的极端情况。一些 Web 服务和 IT 系统，例如 LinkedIn、Salesforce 和其他支持和 HR 应用程序，“合法地冒充”员工发送通知，并更改回复字段以确保对消息的回复被他们的系统记录下来。

其他标头属性可能有助于检测 BEC 攻击。例如，如果电子邮件在一天中的异常时间发送，或者来自异常 IP 或来自国外。然而，许多 BEC 攻击被设计成看似合法的，并且在一天中的正常时间从看似合法的电子邮件地址发送。

身体属性。例子的主体1 包含两个独特的语义属性。首先，它讨论了敏感信息（电汇）。其次，它要求一个特殊的、即时的请求。同样，示例的文本2 正在询问收件人是否可以接受紧急请求。这种对敏感信息或可用性的紧急请求在某些情况下可能是合法的（例如，在财务团队内部的紧急沟通中）。

Example 主体中的唯一属性3 是链接本身。该链接指向一个与公司没有任何关系的网站：它不属于公司通常使用的网络服务，并且与公司的域无关。

最后，所有三个示例都包含特定的文本和视觉元素，这些元素对于发件人的身份而言是独一无二的。例如，例子1 包含 CEO 的签名，所有电子邮件都包含特定的语法和写作

[2]许多组织都有代表他们发送电子邮件的合法系统，例如营销自动化系统，如果电子邮件身份验证设置不当，这些系统可能会被错误地阻止。

风格。如果这些元素中的任何一个偏离了来自特定发件人的普通电子邮件的风格，它们就可以被用来检测假冒行为。由于在许多 BEC 电子邮件中，攻击者非常小心地使电子邮件看起来合法，因此我们不能过分依赖检测文体异常。

如上所示，每个示例都具有独特的异常属性，可用于将其归类为 BEC 攻击。然而，正如我们将在 § 7，这些属性本身都不足以对具有令人满意的误报率的电子邮件进行分类。

利用历史电子邮件。检测电子邮件传播的威胁的许多先前工作都依赖于检测电子邮件中的恶意信号，例如发件人和链接域信誉 [2, 48]，恶意附件 [49]，以及依赖链接点击日志和 IP 登录 [20].然而，作为表1 我们调查的示例表明，大多数 BEC 攻击不包含任何明显的恶意附件或链接。直观地说，访问组织的历史电子邮件将使监督学习系统能够通过识别标题和正文属性中的异常来识别常见的 BEC 攻击类型。我们观察到流行的基于云的电子邮件提供商（例如 Office 365 和 Gmail）使他们的客户能够允许第三方应用程序通过公共 API 以特定权限访问他们的帐户。特别是，这些 API 可以使第三方应用程序能够访问历史电子邮件。这使我们能够设计一个系统，该系统使用历史电子邮件来识别 BEC 攻击。

4 分类器和特征设计

在本节中，我们描述了 BEC-Guard 的设计目标及其训练数据集。然后我们描述了我们在 BEC-Guard 中使用的初始分类器集，并介绍了我们的训练和标记方法。

4.1 设计目标

BEC-Guard 的目标是实时检测 BEC 攻击，而无需系统用户利用安全分析师手动筛选可疑攻击。为了实现这个目标，我们需要优化两个指标：误报率和精度。误报率是误报率占收到的电子邮件总数的百分比。如果我们假设一个普通用户每天收到超过 100 封电子邮件，在一个拥有 10,000 名员工的组织中，我们的目标是很少遇到误报（例如，整个组织每天一次）。因此，我们的目标误报率低于百万分之一。精度是真阳性率（正确检测到 BEC 攻击）占系统检测到的攻击的百分比，而误报率是所有电子邮件（不仅仅是系统检测到的电子邮件）的误报率。如果精度不高，BEC-Guard 的用户将对其预测的有效性失去信心。除了这两个指标外，我们还需要确保高覆盖率，即系统捕获大量大多数 BEC 攻击。

4.2 数据集和隐私

我们使用来自 1,500 个组织的企业电子邮件数据集开发了 BEC-Guard 的初始版本，这些组织是 Barracuda Networks 的积极付费客户。我们数据集中的组织在类型和规模上差异很大。这些组织包括来自不同行业（医疗保健、能源、金融、交通、媒体、教育等）的公司。组织的规模从 10 个邮箱到超过 100,000 个不等。总体而言，为了训练 BEC-Guard，我们标记了 7,000 多个 BEC 攻击示例，这些示例是从 1,500 个组织中随机选择的。

为了访问数据，这些组织授予我们访问其 Office 365 电子邮件环境 API 的权限。 API 提供对所有历史公司电子邮件的访问。这包括在组织内部发送的电子邮件，以及来自所有文件夹（收件箱、已发送、垃圾邮件等）的电子邮件。 API 还允许我们确定每个组织拥有哪些域，甚至是否阅读了电子邮件。

道德和隐私方面的考虑。 BEC-Guard 是商业产品的一部分，参与数据集的 1,500 名客户向 Barracuda Networks 提供了法律许可，允许梭子鱼网络出于识别 BEC 的目的访问其历史公司电子邮件。客户还可以选择随时撤销对 BEC-Guard 的访问权限。

由于数据集的敏感性，在严格的访问控制策略下，它只暴露给开发 BEC-Guard 的五名研究人员。研究团队仅出于标记数据的目的访问历史电子邮件，以开发 BEC-Guard 的分类器。开发分类器后，我们永久删除了所有未主动用于训练分类器的电子邮件。用于分类的电子邮件是加密存储的，只有研究团队可以访问它们。

4.3 将分类分为两部分 BEC 攻击的相对罕见发生影响了我们的几个设计选择。我们的第一个设计选择是排除无监督学习。无监督学习通常使用聚类算法（例如，k-means [15]）对电子邮件类别进行分组，例如 BEC 电子邮件。然而，聚类算法通常会对许多常见类别（例如，社交电子邮件、营销电子邮件）进行分类，但由于 BEC 非常罕见，因此会导致精度低和误报率高。因此，监督学习算法更适合高精度检测 BEC。然而，使用监督学习会带来一系列挑战。

特别是，BEC 是数据不平衡的极端情况。当统一采样时，在我们的数据集中，“合法”电子邮件出现的可能性比 BEC 电子邮件高 50,000。这带来了两个挑战。首先，为了标记适度数量的 BEC 电子邮件（例如 1,000），我们需要标记一个语料库

大约 5000 万封合法电子邮件。其次，即使有大量带标签的电子邮件，训练一个受监督的
众所周知，对不平衡数据集的分类器会导致各种问题，包括使分类器偏爱较大的类别（即合法电子邮件）[24, 26, 47, 51]. 为了处理这种数据不平衡的极端情况，我们将分类和标签问题分为两部分。第一个分类器仅查看电子邮件的元数据，而第二个分类器仅检查电子邮件的正文和主题。

第一个分类器查找冒充电子邮件。我们将冒充定义为以某人的姓名发送的电子邮件，但实际上并非由该人发送。冒充电子邮件包括恶意 BEC 攻击，它们还包括合法冒充员工的电子邮件，例如代表员工发送自动电子邮件的内部系统。模拟分类器仅分析电子邮件的元数据（即发件人、收件人、CC、BCC 字段）。模拟分类器检测到欺骗名称（示例1 和3）和欺骗性电子邮件（示例2）. 第二组分类器，即内容分类器，通过检查电子邮件的主题和正文以查找异常，仅将检测到的电子邮件分类为假冒电子邮件。我们使用两个不同的内容分类器，每个分类器寻找不同类型的 BEC 攻击。[3] 两个内容分类器是：文本分类器，它依靠自然语言处理来分析电子邮件的文本，以及链接分类器，它对可能出现在电子邮件中的任何链接进行分类。

我们所有的分类器都在同一数据集上进行全局训练。然而，为了计算某些特征（例如，发件人姓名和电子邮件地址一起出现的次数），我们依赖于每个组织独有的统计数据。

4.4 模拟分类器

桌子3 包括模拟分类器使用的主要特征。这些特征描述了特定电子邮件地址和名称之前出现在发件人和回复字段中的次数，以及有关发件人身份的统计信息。

为了说明为什么维护特定组织的历史统计数据是有帮助的，请考虑图1. 该图描述了三个月内拥有 44,000 个邮箱的组织中每个发件人使用的电子邮件地址的数量。82% 的用户只从一个地址发送电子邮件，其余的用户从多个地址发送电子邮件。一些发件人使用大量电子邮件地址的原因是他们在 BEC 攻击中反复被冒充。例如，CEO 是一个常见的冒充目标。并且经常成为目标数十次。然而，仅此信号并不

[3]在误报率或准确率方面，使用多个内容分类器并没有固有的优势。我们决定使用两个不同的内容分类器，因为这样可以更方便地分别调试和维护它们。

特征	描述
发件人有公司域？	发件人地址来自公司域吗？
回复！=发件人地址？	回复地址和发件人地址不同？
发件人和电子邮件的次数	发件人姓名和电子邮件地址出现的次数
回复地址次数	回复地址出现次数
已知的回复服务？	回复来自已知的网络服务（例如 LinkedIn）吗？
发件人姓名流行度	发件人姓名有多受欢迎

表 3：模拟分类器使用的主要功能，用于查找模拟尝试，包括欺骗性名称和电子邮件。

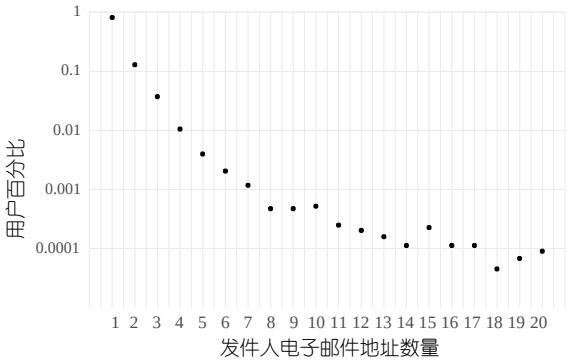

图 1：在拥有 44,400 个邮箱的组织中为每个用户观察到的唯一电子邮件地址数。 X 轴是观察到的唯一电子邮件地址的数量，占组织用户总数的百分比（在 Y 轴中）。

足以检测冒充，。例如，一些拥有大量电子邮件地址的发件人代表共享邮箱（例如，“IT”或“HR”），并且是合法的。

因此，模拟分类器中的一些特征依赖于组织的历史通信模式。这影响了 BEC-Guard 的架构。此外，我们维护了一份已知网络服务的列表，这些服务“合法地”发送回复地址与发件人地址不同的电子邮件（例如 LinkedIn、Salesforce），以便捕获回复。常用服务的原始列表是从主要 Web 服务的域列表中填充的。然后，当我们在标记过程中遇到这些服务时，我们用额外的服务扩充了这个列表（在 § 6 我们讨论了与此合法回复发件人列表相关的可能规避技术）。发件人姓名流行度分数是通过维护一个列表来离线计算的，该列表列出了姓名在我们数据集中不同组织中出现的频率。名字越受欢迎，员工通常不使用的带有电子邮件地址的名字是另一个人的可能性就越大（名字冲突）。

名字和昵称匹配。为了检测名称欺骗，模拟分类器需要匹配

发件人姓名与员工姓名。但是，名称可以写成各种形式。例如："Jane Smith"可以写成："Smith, Jane"、"Jane J. Smith"或"Jane Jones Smith"。此外，我们需要处理名称中可能出现的特殊字符，例如ï或ä。

为了解决这些问题，BEC-Guard 规范化了名称。它将员工姓名存储为〈first name, last name〉元组，并检查发件人姓名的所有变体以查看它是否与具有公司电子邮件地址的员工姓名相匹配。这些变体包括去除中间名或首字母、颠倒名字和姓氏的顺序以及去除后缀。后缀包括像"Jr."这样的例子。或者当电子邮件地址作为发件人姓名的一部分发送时。此外，我们将名字与公开的昵称列表进行匹配[36]，例如当攻击者以"Bill Clinton"的身份发送电子邮件，而员工的姓名被存储为"William Clinton"时，可以捕获这种情况。

内容分类器。我们的系统使用两个内容分类器：文本分类器和链接分类器。文本分类器捕获类似于 Example 的攻击1 和2，并且链接分类器停止了类似于 Example 的攻击3.按照设计，内容分类器比模拟分类器更频繁地更新，并且应该很容易根据用户报告的误报和误报进行重新训练。

文本分类器。在 BEC 攻击中类似于 Example1 和2，正文中包含表示敏感或特殊请求的词语，例如"电汇"或"紧急"。因此，我们的文本分类器的第一次迭代旨在寻找可能暗示特殊请求或财务或 HR 交易的特定词。分类器的特征描述了一系列敏感词和短语在文本中的位置。然而，随着时间的推移，我们注意到这种方法存在一些问题。首先，当攻击者稍微改变特定单词或短语时，依赖于硬编码关键字的分类器可能会错过攻击。其次，为了成功地重新训练分类器，我们必须修改它寻找的关键字列表，这需要每天手动更新关键字列表。

相反，我们开发了一个文本分类器，它可以学习自己指示 BEC 的表达式。第一步是预处理文本。 BEC-Guard 从电子邮件的主题和正文中删除对电子邮件分类无用的信息。它删除了正则表达式模式，包括称呼（"Dear"、"Hi"）、预制标题以及页脚（"Best"）和签名。它还会删除所有英文停用词，以及可能出现在电子邮件中的任何名称。

第二步是计算频率逆文档频率[42]（TFIDF）电子邮件中每个单词的分数。 TFIDF 表示每个单词在电子邮件中的重要性，定义为：

$$TF(w) = \frac{w \text{ 在电子邮件中出现的次数}}{\text{电子邮件中的字数}}$$

$$国防军(w) = \frac{\text{日志（电子邮件数量）}}{\text{带 w 的电子邮件数量}}$$

其中 w 是电子邮件中的给定单词。 TF(w)· IDF(w) 对在特定电子邮件中频繁出现但在整个电子邮件语料库中相对罕见的单词给予更高的分数。直觉是，在 BEC 电子邮件中，例如表示紧急或特殊请求的词将具有较高的 TFIDF 分数，因为它们经常出现在 BEC 电子邮件中，但在合法电子邮件中出现频率较低。

在训练文本分类器时，我们计算训练集中每封电子邮件中每个单词的 TFIDF 分数。我们还计算词对（二元组）的 TFIDF。我们将 IDF 的全局统计信息存储为字典，其中包含训练集中的电子邮件数量，这些电子邮件包含在文本分类器训练中遇到的独特短语。我们将字典大小限制为 10,000 个排名靠前的单词（我们在§中评估了字典大小如何影响分类精度7.2).

每封邮件的特征向量等于字典中单词的个数，每个数字代表字典中每个单词的TFIDF。未出现在电子邮件中或未出现在字典中的单词的 TFIDF 为零。最后一步是根据这些特征运行分类器。桌子4 展示了我们数据集中 BEC 电子邮件中的前 10 个短语（unigram 和 bigram）。请注意，最上面的短语都表示某种形式的紧迫感。

TFIDF 的 BEC 电子邮件中的热门短语	
1. 有时间	6. 需要完成
2. 回复	7. 尽快
3. 时刻需要	8. 紧急响应
4. 片刻	9. 紧急
5. 需要	10. 完成任务

表 4：BEC 电子邮件的前 10 个短语，按它们在我们的评估数据集中的 TFIDF 排名排序（有关评估数据集的更多信息，请参阅§7.1）. TFIDF 是针对我们评估数据集中所有 BEC 电子邮件中的每个单词计算的。

链接分类器。链接分类器检测到类似于 Example 的攻击3.在这些攻击中，攻击者试图让收件人点击钓鱼链接。正如我们之前所述，这些个性化的网络钓鱼链接通常不会被 IP 黑名单检测到，并且通常对收件人来说是独一无二的。在这种情况下，由于内容分类器仅将已分类为冒充电子邮件的电子邮件分类，因此它可以将链接标记为"可疑"，即使它们的误报率很高。例如，指向一个链接

小型网站或最近注册的网站，结合假冒尝试很可能是 BEC 电子邮件。

特征	描述
域名流行度	链接的最不受欢迎的域有多受欢迎
网址字段长度	最不受欢迎的 URL 的长度（长 URL 比较可疑）
域名注册日期	最少人口的域名注册日期 lar 域名（新域名可疑）

表 5：链接请求分类器使用的主要功能，它阻止了示例中的攻击3.

桌子5 描述了链接请求分类器使用的主要特征。域流行度是通过测量域的 Alexa 分数来计算的。为了处理链接缩短器或链接重定向，BEC-Guard 在为链接分类器计算它们的特征之前扩展 URL。此外，一些 URL 特征需要确定有关域的信息（流行度和分数）。对于域名流行度特征，我们缓存了一个最流行的域名列表，并离线更新它。为确定域名注册日期，BEC-Guard 会进行实时 WHOIS 查询。请注意，与需要映射每个发件人姓名的电子邮件地址分布的模拟分类器不同，文本和链接分类器的特征都不是特定于组织的。这使我们能够根据用户报告的电子邮件轻松地重新训练它们。

4.5 分类器算法

模拟和链接分类器使用随机森林 [5] 分类。随机森林由随机形成的决策树组成 [40]，其中每棵树投一票，决定由大多数树决定。我们的系统使用随机森林而不是单个决策树，因为我们发现它们提供了更好的精度，但是对于离线调试和分析，我们经常可视化单个决策树。我们决定将 KNN 用于文本分类器，因为它的覆盖率略高于随机森林。然而，我们发现由于文本分类器使用了大量的特征（一个包含 10,000 个短语的字典），它在不同分类器中的功效是相似的。在 §7.2 我们评估不同分类器算法的性能。此外，我们还探索了基于深度学习的技术，例如 word2vec [34] 和 sense2vec [46]，它将每个单词扩展为表示其不同含义的向量。我们目前不使用这种深度学习技术，因为它们在训练和在线分类方面的计算量很大。

检测新员工的冒充行为。当新员工加入组织时，模拟分类器将没有足够的历史信息

员工，因为他们不会有任何历史电子邮件。随着该员工收到更多电子邮件，BEC-Guard 将开始为该员工编制统计数据。定期清除旧电子邮件的组织也可能出现类似的问题。在实践中，我们发现分类器仅在一个月的数据后就表现良好。

4.6 标签

为了标记初始训练集，我们对 BEC 攻击模型做了几个假设。首先，我们假设攻击者使用他们的名字冒充员工（在一组允许的变体下，如上所述）。其次，我们假设使用同一个电子邮件地址进行模拟的次数不会超过 100 次。第三，我们假设攻击者使用不同于公司地址的电子邮件地址，作为发件人地址或回复地址。我们讨论不符合这些假设的其他类型的攻击，以及攻击者如何在 §中规避这些假设6. 在这些限制下，我们完全涵盖了所有可能的攻击并手动标记它们。此外，我们还纳入了客户报告的未命中攻击（我们在§中讨论了这个过程7.3).

我们假设 BEC 电子邮件不会冒充员工使用同一电子邮件地址超过 100 次的原因是 BEC-Guard 的设计假设该组织已经在使用垃圾邮件过滤器，它可以防止基于容量的攻击（例如，Office 365 或 Gmail 的默认垃圾邮件保护）。因此，从未知地址向同一收件人发送电子邮件超过 100 次的攻击者很可能会被垃圾邮件过滤器阻止。事实上，在我们仅由垃圾邮件过滤后的电子邮件组成的整个数据集中，我们从未见过攻击者使用电子邮件地址冒充员工超过 20 次。请注意，我们仅将此假设用于标记原始训练集，而不将其用于正在进行的再训练（因为再训练是基于客户报告的攻击）。

模拟分类器。为了标记模拟分类器的训练数据，我们对原始电子邮件的标头运行查询，以发现在我们的标记假设下可能包含 BEC 攻击的所有电子邮件（见上文）。然后，我们将所有查询电子邮件的结果标记为冒充电子邮件，并将查询未找到的所有电子邮件标记为合法电子邮件。

内容分类器。内容分类器的训练数据集是通过在新数据集上运行经过训练的模拟分类器构建的，然后对其进行手动标记。我们用于内容分类器的初始训练集包括 300,000 封来自随机选择的组织超过一年的模拟电子邮件数据。即使在这个训练数据集中，我们也能够进一步显着限制需要手动标记的电子邮件数量。这是因为绝大多数这些电子邮件显然是

不是 BEC 攻击，因为它们是由冒充大量员工的合法 Web 服务引起的（例如，帮助台系统代表 IT 员工发送电子邮件）。

构建初始数据集后，训练内容分类器非常简单，因为我们不断地从用户那里收集假阴性和假阳性电子邮件并将它们添加到训练集中。请注意，我们仍然在重新训练之前手动检查这些样本作为质量控制措施，以确保对手不会“毒害”我们的训练集，并确保用户没有错误地标记电子邮件。

对数据集进行采样。在不平衡的数据集上天真地训练分类器通常会使分类器偏向于多数类。具体来说，它可以导致分类器总是简单地选择预测多数类别，即合法电子邮件，从而实现非常高的准确性（即准确性 = (t p + tn)/(t p + tn + f p + f n)，其中 t p 是真阳性，tn 是真阴性，f p 是假阳性，f n 是假阴性）。由于 BEC 在我们的数据集中非常罕见，因此始终预测电子邮件合法的分类器将实现高精度。这个问题在我们的模拟分类器的情况下尤为严重，它需要在合法电子邮件和 BEC 电子邮件之间进行初始过滤。对于内容分类器，我们不必对数据集进行采样，因为它处理的训练数据集要小得多。有多种处理不平衡数据集的方法， 包括对少数类进行过采样和对多数类进行欠采样[6, 24, 27, 29, 30]，以及将更高的成本分配给错误预测少数类[9, 38].

我们的第二个主要设计选择是对多数类别（合法电子邮件）进行欠采样。我们做出这个决定有两个原因。首先，如果我们决定对 BEC 进行过度采样

攻击，我们需要在很大程度上这样做。这可能会使我们的分类器过度拟合，并基于相对较少的正样本使结果产生偏差。其次，过度采样使训练在计算上更加昂贵。

一种天真的欠采样方法是对合法电子邮件进行统一采样。然而，这会导致分类器精度较低，因为不同类别的合法电子邮件没有得到很好的表示。例如，对电子邮件进行统一抽样可能会漏掉来自合法冒充员工的 Web 服务的电子邮件。模拟分类器会将这些电子邮件标记为 BEC 攻击，因为它们在训练数据集中相对较少。

对多数类进行欠采样的主要挑战是如何用相对较少的样本（即与 BEC 电子邮件样本的数量相当或相等）来表示整个合法电子邮件的宇宙。为此，我们使用无监督学习算法高斯混合模型（GMM）对合法电子邮件进行聚类。聚类算法将样本分成簇，每个簇由正态分布表示，投影到模拟分类器特征空间。数字2 说明了一个

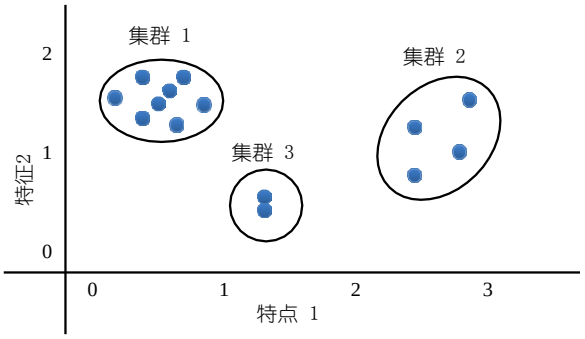

图 2：在具有三个集群的二维特征空间中对一组合法电子邮件运行聚类算法的描述。在对合法电子邮件进行聚类后，我们根据集群的大小从每个集群中选择样本数量。

具有两个特征和 14 个合法电子邮件样本的示例。在这个例子中，样本被分成三个集群。为了选择合法电子邮件的代表性样本，我们从每个集群中随机选择一定数量的样本，与属于每个集群的合法电子邮件数量成正比。例如，如果我们的目标是总共使用 7 个样本，我们将从第一个簇中选择 4 个样本，从第二个簇中选择 2 个样本，从第三个簇中选择 1 个样本，因为每个簇中的原始样本数是 8 、4 和 2。

我们选择的集群数量保证合法电子邮件的每个主要“类别”的最小表示。我们发现使用 85 个集群足以捕获我们数据集中的合法电子邮件。当我们尝试使用超过 85 个集群时，第 85 个集群之后的集群几乎或完全是空的。即使在对模拟分类器进行多次迭代重新训练之后，我们发现 85 个集群足以表示我们的数据集。

5 系统设计

BEC-Guard 由两个关键阶段组成：在线分类阶段和离线训练阶段。定期（每隔几天）进行离线培训。当一封新电子邮件到达时，BEC-Guard 会结合模拟和内容分类器来确定电子邮件是否为 BEC。这些分类器在离线训练阶段提前训练。我们在下面更详细地描述了我们系统设计的关键组件。

传统上，商业电子邮件安全解决方案具有网关架构，或者换句话说，它们位于入站电子邮件的数据路径中并过滤恶意电子邮件。如上所述，BEC-Guard 的某些模拟分类器功能依赖于内部通信的历史统计数据。网关架构对检测 BEC 攻击施加了限制，原因有二。首先，网关通常无法观察内部通信。其次，网关通常无法访问历史通信，因此需要几个月或更长时间的通信模式观察才能让系统检测到

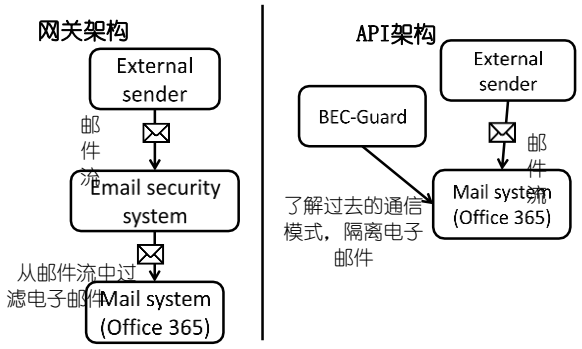

图 3： 传统电子邮件安全系统的架构与 BEC-Guard 架构之间的比较，传统电子邮件安全系统作为网关在电子邮件到达邮件系统之前对其进行过滤，BEC-Guard 架构依赖于 API 来了解每个组织的历史通信模式，以及实时检测攻击。

传入的 BEC 攻击。幸运的是，基于云的电子邮件服务（例如 Office 365 和 Gmail）提供的 API 可以访问历史通信以及实时监控和移动电子邮件。 BEC-Guard 利用这些 API 既可以访问历史通信，也可以进行近乎实时的 BEC 检测。数字3 将网关架构与 BEC-Guard 基于 API 的架构进行比较。我们使用 Office 365 API 描述 BEC-Guard 的设计和实施。

热身阶段。我们将分析每个组织的历史通信的过程命名为热身阶段。为了开始预热，该组织允许 BECGuard 通过使用 OAuth 和 Office 365 管理员帐户的身份验证令牌访问其 Office 365 帐户。这允许 BEC-Guard 访问与该帐户关联的所有用户的 API。一旦通过身份验证，BEC-Guard 就开始收集模拟分类器所需的统计数据（例如，某个用户从某个电子邮件地址发送电子邮件的次数）。 BEC-Guard 收集的统计数据可以追溯到一年前。我们发现分类器仅用一个月的历史数据就表现良好。

在线分类。在预热阶段之后，BECGuard 准备好实时检测传入的 BEC 攻击。为此，BEC-Guard 等待来自组织 Office 365 帐户中任何用户的 Webhook API 调用。只要特定用户有任何新活动，webhook API 就会调用 BEC-Guard。当 webhook 被触发时，BEC-Guard 检查是否有新收到的电子邮件。如果是这样，BEC-Guard 检索电子邮件并对其进行分类，首先使用模拟分类器，使用包含每个组织唯一的历史通信统计数据的数据库。然后，只有当它被归类为模拟电子邮件时，BEC-Guard 才会使用内容分类器对电子邮件进行分类。

如果至少一个内容分类器将电子邮件归类为 BEC 攻击，则 BEC-Guard 会隔离该电子邮件。这是通过从用户接收电子邮件的文件夹（通常是收件箱文件夹）中删除电子邮件并将其移动到最终用户的指定隔离文件夹中来执行的邮箱。由于电子邮件在服务器端被隔离，当用户的电子邮件客户端同步电子邮件时，它也会在用户的电子邮件客户端上被隔离。此外，绝大多数电子邮件在同步到用户的电子邮件客户端之前都会被 BEC-Guard 隔离。

6 逃避

在本节中，我们将讨论 BEC-Guard 当前未阻止的攻击，以及攻击者可以用来绕过 BEC-Guard 的规避技术以及如何解决这些问题。 BEC-Guard 是生产中的实时服务，并且已经发展

自 2017 年首次推出以来，它迅速发展。我们已经部署了额外的分类器来增强本文中描述的分类器，以响应下面介绍的一些规避技术，并且现有的分类器已经过多次重新训练。基于 API 的架构的另一个好处是，如果我们发现一些攻击被规避漏掉了，我们可以及时回溯并找到它们，并相应地更新系统。电子邮件威胁形势瞬息万变，虽然检测器保持高精度很重要，但安全系统可以轻松调整和重新培训也同样重要。

6.1 阻止其他攻击

BEC-Guard 专注于阻止 BEC 攻击，在这种攻击中，外部攻击者冒充员工。但是，BEC-Guard 未涵盖其他类型的 BEC。帐户接管。当攻击者窃取员工的凭据时，他们可以远程登录以向其他员工发送 BEC 电子邮件。我们将此用例称为“帐户接管”。有几种检测帐户接管的方法，包括监控内部电子邮件的异常情况（例如，一名员工突然向他们通常不与之通信的其他员工发送许多电子邮件）、监控可疑的 IP 登录以及监控可疑的收件箱规则更改（例如，员工突然创建删除出站电子邮件的规则）[18–20].这种场景不是BEC-Guard的重点，但是我们的商业产品已经涵盖了。

在不更改回复地址的情况下冒充发件人姓名和电子邮件。外部攻击者可能会发送冒充发件人姓名和电子邮件地址的电子邮件，而无需使用不同的回复地址。我们没有在我们的数据集中观察到此类攻击，但它们是可能的，尤其是在攻击者要求收件人点击链接以窃取其凭据的情况下。与帐户接管类似，可以通过查找异常电子邮件模式来检测此类攻击。 Gascon 等人使用的另一种可能的方法是在实际的 MIME 标头中寻找异常 [14].

冒充外部人员。 BEC-Guard 的模拟分类器目前依赖于访问员工的历史入站电子邮件。为了检测频繁通信的外部人员的冒充

对于组织，BEC-Guard 可以合并从外部人员发送到公司的电子邮件。

任何语言的文本分类。 BEC-Guard 目前经过优化，可以捕获我们数据集中频繁出现的语言的 BEC。模拟分类器和链接分类器均不依赖于语言，但文本分类器依赖于 TFIDF 词典，依赖于标记数据集的语言。有几种可能的方法可以使 BEC-Guard 的文本分类器完全与语言无关。一种是有意收集足够多的各种语言样本（基于用户报告或综合生成），并对这些电子邮件进行标记和训练。另一种可能更具可扩展性的方法是翻译带标签的电子邮件（例如，使用谷歌翻译或类似工具）。通用发件人姓名。BEC-Guard 明确尝试检测假冒员工姓名。但是，攻击者可能会冒充更通用的名称，例如“HR 团队”或“IT”。这种攻击超出了本文的范围， 但我们使用与 BEC-Guard 类似的方法来解决它以检测这些攻击：我们将我们的内容分类器与新的模拟分类器相结合，该分类器查找通常出现在不同组织中但从非公司发送的发件人姓名电子邮件地址或具有不同的回复地址。

品牌模仿。与“通用发件人”攻击类似，攻击者经常冒充流行的在线服务（例如，Google Drive 或 Docusign）。这些类型的攻击超出了本文的范围，但我们使用类似的方法检测它们，将内容分类器与寻找异常发件人的模拟分类器（例如，发件人姓名为“Docusign”，但发件人域与 Docusign 无关）。

6.2 逃避检测

除了 BEC-Guard 并非旨在检测的 BEC 攻击（如上所述）之外，攻击者还可以通过其他几种方式来尝试规避 BEC-Guard。我们将在下面讨论这些问题，并讨论我们如何调整 BEC-Guard 来解决这些问题。

使发件人电子邮件地址合法化。任何使用基于异常检测的信号的系统都容易受到攻击者的攻击，这些攻击者会付出额外的努力来避免出现“异常”。例如，在标记我们的数据集时，我们假设被模拟的员工被同一发件人电子邮件地址模拟的次数不超过 100 次。虽然这个阈值没有硬编码到模拟分类器中，但它是我们用来为初始训练集过滤电子邮件的阈值，因此可能会使分类器产生偏差。请注意，我们从未观察到攻击者使用同一电子邮件冒充员工超过 20 次。

我们相信这个假设是有效的，因为 BEC-Guard 假设组织已经在使用基于卷的安全过滤器（例如，O365 的默认垃圾邮件保护或 Gmail 或其他垃圾邮件过滤器），这会引起“容量”攻击。通常，这些系统会将一封从未知地址一次性发送给 100 多名员工的电子邮件标记为垃圾邮件。

但是，老练的攻击者可能会尝试通过从模拟电子邮件地址向特定组织发送大量合法电子邮件来绕过这些过滤器，并且只有在发送数百封合法电子邮件后，他们才会使用该地址发送 BEC。当然，这种方法的缺点是它需要攻击者进行更多投资，并增加执行成功的 BEC 活动的经济成本。克服这种攻击的一种方法是将人工样本添加到具有更高阈值的模拟分类器中，以消除偏差。当然，这可能会降低 BEC-Guard 的整体精度。

使用不常见的同义词。另一种规避技术是发送包含与用于训练我们的文本分类器的标记电子邮件不同或 TFIDF 较低的文本的电子邮件。例如，“银行”一词的 TFIDF 高于“基金”一词。如前所述，克服这些类型攻击的一种方法是使用一种技术覆盖同义词，例如 word2vec [34].

操纵字体。攻击者采用了各种字体操作来避开基于文本的检测器。例如，一种技术是使用大小为零 [35]]，它们不会显示给最终用户，但可用于混淆文本的模拟或含义。另一种技术是使用非拉丁字母，例如西里尔字母，它们对最终用户来说看起来类似于拉丁字母，但不会被基于文本的检测器解释为拉丁字母 [16].

为了处理这些类型的技术，我们总是在将任何文本提供给 BEC-Guard 的分类器之前对其进行规范化。例如，我们忽略任何字体大小为零的文本。如果我们遇到西里尔字母或希腊字母与拉丁文本结合使用，我们会规范化非拉丁字母以匹配外观最接近它的拉丁字母。虽然这些技术是基于启发式的，但它们已被证明可以有效阻止常见形式的基于字体的规避。

在图像中隐藏文本。攻击者可以将文本隐藏在嵌入的图像中，而不是在电子邮件中使用文本。我们在实践中很少观察到这种用例，很可能是因为这些攻击可能不太有效。默认情况下，许多电子邮件客户端不显示图像，即使显示图像，电子邮件对收件人来说也可能看起来很奇怪。因此，我们目前不解决这个用例，但解决它的一种直接方法是使用 OCR 提取图像中的文本。

使用合法的回复地址。如§所述4.4 BEC-Guard 依靠合法回复域列表来减少误报。此列表可能会被利用。例如，攻击者可以制作与被模拟员工同名的 LinkedIn 或 Salesforce 个人资料，并从该服务发送模拟电子邮件。

	精确	计划生育	记起
BEC-Guard（组合）仅模拟	98.2%	0.000019%	96.9%
	11.7%	（5,260,000 分之一）	100%
		0.016%	

表 6：与单独的模拟分类器相比，BEC-Guard 的精确度、误报率和召回率。

虽然这确实是一种潜在的规避技术，但这些第三方服务通常有自己的反欺诈机制来阻止假冒。此外，我们认为如果通过第三方服务进行假冒尝试，则不太可能成功，因为它可能看起来比简单地从员工的电子邮件帐户发送电子邮件要自然得多。无论如何，我们从未见过攻击者使用这种规避技术。

7 评估

在本节中，我们评估 BEC-Guard 的功效。我们首先结合模拟和内容分类器分析 BEC-Guard 的端到端性能。然后我们分解每组分类器的性能，分析不同分类器算法的性能。我们还尝试通过比较客户报告的未命中攻击的数量与真阳性的数量来估计未被 BEC-Guard 捕获的未知攻击的程度。

7.1 端到端评估

对于端到端评估，我们随机抽取了 2018 年 6 月由 BEC-Guard 处理的电子邮件。我们手动标记了电子邮件，并在标记数据上评估了 BEC-Guard 的分类器。我们为评估数据集标记的电子邮件类似于我们为 BEC-Guard 的分类器标记训练数据的方式（见 § 4.6）. 我们首先运行了一组查询，以发现我们在标签假设下可以找到的所有 BEC 攻击。然后我们手动标记生成的电子邮件，并找到 4,221 封 BEC 电子邮件。整个过程需要一个人大约一周的工作时间。未标记为 BEC 攻击的电子邮件被认为是无辜的（在 § 7.3 我们讨论了我们的标记过程可能遗漏的电子邮件）。

为了评估分类器，我们将评估数据集随机分成两半：我们使用一半的电子邮件进行训练，其余的用于测试分类器。该数据集包括来自数百个组织的 2 亿封电子邮件。

为了测试 BEC-Guard 的端到端功效，我们仅对被模拟分类器检测为模拟电子邮件的电子邮件运行内容分类器。桌子6 总结疗效结果。在我们标记的电子邮件中，BEC-Guard 的召回率很高：我们标记的 BEC 电子邮件中有 96.9% 被模拟分类器和其中一个内容分类器成功分类。综合误报率仅为 530 万封电子邮件中的一封

文本分类器			
算法	精确	计划生育	记起
逻辑回归	97.1%	$6.1{\cdot}10^{-5}$%	98.4%
线性支持向量机	98.3%	$3.6{\cdot}10^{-5}$%	98.7%
决策树	96.0%	$8.5{\cdot}10^{-5}$%	97.1%
随机森林	99.2%	$1.7{\cdot}10^{-5}$%	96.4%
韩国国家网络	98.9%	$2.3{\cdot}10^{-5}$%	97.5%

表 7：使用包含 10,000 个单词的字典的文本分类器算法效果。文本分类器算法的功效之间几乎没有区别。

链接分类器			
算法	精确	计划生育	记起
逻辑回归	33.3%	$85.7{\cdot}10^{-5}$%	96.0%
线性支持向量机	92.3%	$3.2{\cdot}10^{-5}$%	90.8%
决策树	94.9%	$2.3{\cdot}10^{-5}$%	96.3%
随机森林	97.1%	$1.3{\cdot}10^{-5}$%	96.0%
韩国国家网络	92.5%	$3.3{\cdot}10^{-5}$%	93.5%

表 8：链接分类器算法功效。随机森林提供优于其他算法的结果。

错误检测，这超出了我们百万分之一电子邮件的设计目标。准确率为98.2%。

组合分类器的误报是由于模拟分类器检测到的电子邮件（例如，由于个人电子邮件地址）也包含异常内容（例如，员工使用个人电子邮件转发具有低流行域的链接）的不太可能发生的事件给同事）。另一种常见的误报发生在员工离开组织并出于税务目的或其他个人信息要求 W-2 表格时。我们计划通过合并指示发件人是否不再是组织雇员的功能来解决此类误报（例如，如果他们已停止从其公司地址发送电子邮件）。漏报主要是由于 URL 不被视为可疑的情况，因为它属于被入侵的具有相对较高域流行度的域，或者因为电子邮件的文本未被归类为可疑。 后一种情况通常是因为攻击者没有使用类似于用于训练文本分类器的任何 BEC 攻击的短语。例如，其中一个漏报要求收件人提供礼品卡信息，这不是任何先前攻击中使用的请求。

我们还在评估数据集上运行了模拟分类器。其准确率为 11.7%，误报率为 0.016%。只关心召回并有能力容忍相对大量的错误警报的组织可以自行运行模拟分类器。假冒分类器的绝大多数误报是由于员工使用他们的个人或大学（校友）电子邮件地址造成的。

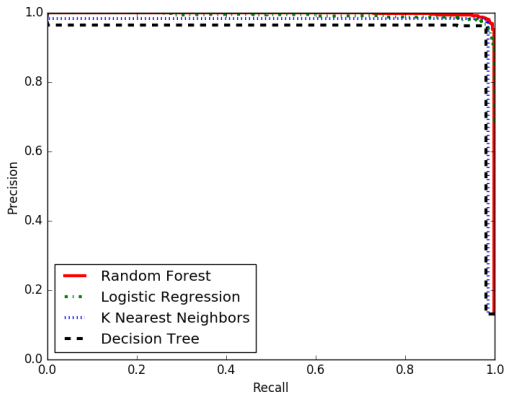

图 4：不同算法的文本分类器的 ROC 曲线。所有四种算法的表现都非常相似，并在大约 99% 的召回率时达到精度悬崖。

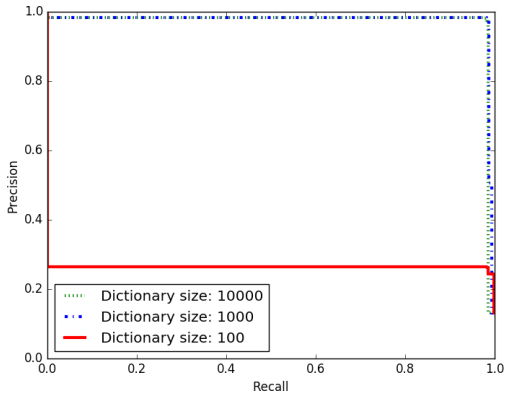

图 5：使用具有不同字典大小的 KNN 的文本分类器的 ROC 曲线。大小为 1,000 的字典已经提供了大部分好处。

7.2 分类器算法

桌子7 比较使用不同分类器算法的文本分类器的结果。正如结果所示，不同分类器之间的差异非常小。这主要是因为我们使用了具有大量特征（10,000）的字典。桌子8 显示链接分类器的结果。在链接分类器的情况下，随机森林比其他分类器（包括 KNN）更明显地提供了更好的结果。链接分类器对分类算法更敏感，因为它使用的特征数量更少。数字4 展示了四种具有概率输出的分类器算法的 ROC 曲线。 ROC 曲线显示了如何调整每个分类器以权衡召回率的精度。所有四种算法的行为几乎相同：它们提供高水平的精确度，直到召回率接近 99%，此时它们的精确度下降。请注意，为了生成 ROC 曲线，我们仅对已分类为模拟的电子邮件运行文本分类器。因此，其在 ROC 曲线中的最小精度等于约 11.7%，这等于

组织	TP	纤维蛋白	原因
A	31	1	通用发件人姓名
B	4	1	错误分类的内容
C	12	1	外部模拟
D	8	1	外部模拟
E	5	1	错误分类的内容
全部的	60		5

表 9：五个组织中的真阳性（TP）和报告的假阴性（FN），其中管理员至少报告了一个假阴性。

模拟分类器。

为了分析字典大小对分类的影响，图5 绘制了使用具有不同字典大小的 KNN 的文本分类器的功效。该图显示，大部分边际收益是在字典大小为 1,000 时实现的。当使用大于 10,000 的字典时，我们没有观察到明显的功效差异。

7.3 评估未命中的攻击

评估不平衡数据集的一个普遍限制是很难准确估计真正的假阴性率。在我们的评估数据集中，我们只能估计与我们标记的数据相关的假阴性率。如果我们在标记过程中错过了一次攻击，并且没有被分类器检测到，我们就不会将其视为漏报。

为了应对“未知”攻击，我们的生产系统允许用户报告它没有检测到的攻击。我们估计了已报告未命中攻击的组织中未命中攻击的数量。我们随机选择了五个报告未命中攻击的组织，并分析了他们在报告未命中攻击的月份的检测结果。桌子9 提供了这五个组织中真实和漏检的数量，以及每个假阴性的原因。

在组织 A 中，攻击被遗漏了，因为电子邮件没有冒充员工姓名，而是发件人姓名有一个通用标题（例如，“会计师”）。 BEC-Guard 仅检测对员工姓名的冒充。正如我们在标签假设中所解释的那样（见 § 4.6），BEC-Guard 仅用于检测明确冒充员工姓名的攻击。我们推测这种类型的电子邮件不太成功，因为收件人可能会发现从发件人姓名中收到带有通用标题的电子邮件很不寻常，这在他们的公司中通常不使用。尽管如此，我们的商业产品也使用其他检测器来查找“通用标题”（参见 § 6）. 在组织 B 和 E 中，模拟分类器成功检测到模拟，但文本分类器并未将电子邮件文本视为可疑。在这两种情况下，我们都使用报告的电子邮件重新训练了 BEC-s 文本分类器。在组织 C 和 D 的情况下，报告的丢失电子邮件是由于外部同事的冒充（例如，与公司合作的供应商被冒充）。在 §6 我们

讨论如何扩展 BEC-Guard 以检测此类攻击。

8 相关工作

BEC 日益增长的威胁是众所周知的，并且在许多行业和政府报告中都有描述 [13, 22, 23]. 然而，现有的学术工作使用非常小的或合成的数据集，并且存在高误报率。此外，由于现有的相关工作是基于有限的数据集，它无法解决我们论文中讨论的许多现实问题，例如处理不平衡的数据集、员工使用个人电子邮件地址或"合法"冒充. 我们认为相关工作很少的原因是 BEC 主要影响企业用户（而非消费者），学术研究人员通常很难获得对企业电子邮件数据的访问权限。

电子邮件档案 [10] 在收到的电子邮件上建立行为模型以阻止 BEC。然而，它仅基于 20 个邮箱，没有真实世界的攻击示例，也没有报告误报率。此外，在检测电子邮件的系统上有先前的工作，这会通过网络钓鱼链接破坏员工凭证 [20,45]. BEC 攻击和破坏凭据的电子邮件之间存在一些重叠：在我们的数据集中，40% 的 BEC 攻击试图通过链接钓鱼员工凭据。但是，其余 BEC 攻击不包含危及凭据的网络钓鱼链接，并且无法被这些系统检测到。

加斯科等人。 [14] 设计一个模型来阻止欺骗收件人域的电子邮件。与 BEC-Guard 类似，他们的模型基于发件人的历史通信模式。然而，在我们的数据集中，欺骗电子邮件仅占 BEC 攻击的 1% 左右。因此，他们的模型不会捕获其他 99% 的 BEC 攻击。域欺骗只占我们数据集一小部分的原因是我们的数据集仅包含已被现有垃圾邮件过滤器（例如 Office 365 的默认过滤器）过滤的电子邮件。域欺骗电子邮件包含发件人域和回复域之间或发件人域与发件人电子邮件信封之间的不匹配。出于这个原因，传统的垃圾邮件过滤器已经阻止了大量的欺骗邮件[33]. 此外，他们的模型基于仅包含 92 个邮箱的数据集。达斯 [20] 使用无监督学习技术来识别导致凭据盗窃的结果，这是 BEC 攻击的一个子集。但是，它无法检测仅包含纯文本的攻击，并且基于来自仅包含 19 种已知攻击的单个组织的数据集。它还具有 0.2% 的精度和比 BEC-Guard 高得多的误报率。同样，IdentityMailer [45] 试图通过模拟员工行为和检测出站电子邮件中的异常来防止员工凭证泄露。一旦检测到异常，员工将被要求使用双因素身份验证重新进行身份验证。然而，他们的技术存在非常高的误报率（1%-8%，而 BEC-Guard 中的误报率为百万分之一），并且分析基于一小部分电子邮件。

Ho 等人在 Barracuda Networks 进行的另一项同期研究。 [18, 19] 检查攻击者使用受损帐户的行为以及检测帐户接管事件的可能方法。本文介绍的技术与其他研究相得益彰，并侧重于不同类型的攻击。

最后，在垃圾邮件检测的背景下，有大量关于逆向学习的工作[3, 4, 8, 21, 31, 32, 37, 50] 这与我们的工作有关。未来，我们计划结合过去工作中引入的一些规避技术，包括随机化和使用蜜罐来欺骗对手。

9 结论

BEC 是一种重大的网络安全威胁，每年造成数十亿美元的损失。我们推出了第一个以高精度和误报检测各种 BEC 攻击的系统，并被数千个组织使用。 BEC-Guard 使用基于 API 的新颖架构与监督学习相结合，实时防止这些攻击。

我们在开发和部署 BEC-Guard 过程中吸取的主要教训之一是，攻击者不断调整他们的策略和方法。虽然我们的监督学习方法确实需要不断地重新训练我们的分类器，并且不能完全推广，但我们发现通过基于 API 的架构使用历史电子邮件模式的一般方法对于快速开发新的分类器以应对不断发展的威胁非常有用。我们在其他情况下采用了与本文中描述的方法类似的方法，例如检测品牌冒充、通用发件人名称和帐户接管。

致谢

我们感谢 Grant Ho、我们的牧羊人 Devdatta Akhawe 和匿名审稿人的深思熟虑的反馈。

参考文献

[1] R 安格伦。 第一次 凤凰城购房者在 2017 年的房地产骗局中受骗 7.3 万美元。https://www.azcentral.com/story/news/local/ 亚利桑那州-调查/2017/12/05/第一- 时间-凤凰-购房者-骗出-73-K- 房地产骗局/667391001/.

[2] Manos Antonakakis、Roberto Perdisci、David Dagon、Wenke Lee 和 Nick Feamster。为 DNS 构建动态信誉系统。在第 19 届 USENIX 安全会议记录中，USENIX Security'10，第 18-18 页，美国加利福尼亚州伯克利，2010 年。USENIX 协会。

[3] Marco Barreno、Blaine Nelson、Anthony D. Joseph 和 J. D. 泰格。机器学习的安全性。机器学习，81(2):121–148，2010 年 11 月。

[4] Marco Barreno、Blaine Nelson、Russell Sears、Anthony D. Joseph 和 J. D. Tygar。机器学习可以安全吗？在 2006 年 ACM 信息、计算机和通信安全研讨会论文集中，ASIACCS '06，第 16-25 页，美国纽约州纽约市，2006 年。ACM。

[5] 里奥·布雷曼。随机森林。机器学习，45(1):5–32，2001 年 10 月。

[6] Nitesh V. Chawla、Kevin W. Bowyer、Lawrence O. Hall 和 W. Philip Kegelmeyer。 Smote：合成少数过采样技术。J. Artif。诠释。研究，16(1):321–357，2002 年 6 月。

[7] A. 西顿。 威胁聚焦：针对抵押贷款的鱼叉式网络钓鱼。钩一个大的.，2017.https://blog.barracuda.com/2017/07/31/威胁聚光灯鱼叉式网络钓鱼-抵押贷款-勾搭一个大一个/.

[8] Nilesh Dalvi、Pedro Domingos、Mausam、Sumit Sanghai 和 Deepak Verma。对抗性分类。在第十届 ACM SIGKDD 知识发现和数据挖掘国际会议记录中，KDD '04，第 99-108 页，美国纽约州纽约市，2004 年。ACM。

[9] 佩德罗·多明戈斯。 Metacost：一种使分类器对成本敏感的通用方法。在第五届 ACM SIGKDD 知识发现和数据挖掘国际会议论文集中，第 155-164 页。美国计算机学会，1999 年。

[10] Sevtap Duman、Kubra Kalkan-Cakmakci、Manuel Egele、William Robertson 和 Engin Kirda。 EmailProfiler：具有电子邮件标题和样式特征的鱼叉式网络钓鱼过滤。计算机软件和应用会议 (COMPSAC)，2016 年 IEEE 第 40 届年会，第 1 卷，第 408-416 页。 IEEE，2016 年。

[11] Luca Invernizzi Elie Bursztein， Kylie McRoberts. 跟踪 桌面 勒索软件 端到端付款。 2017 年美国黑帽大会，2017 年。https://www.elie.net/talk/tracking-desktop-勒索软件支付端到端.

[12] 联邦调查局。 2017 年全球网络金融欺诈呈上升趋势。https://www.fbi.gov/news/stories/商业电子邮件妥协在增加.

[13] 联邦调查局。商业电子邮件泄露，120 亿美元的骗局，2018 年 。 https://www.ic3.gov/media/2018/180712.aspx.

[14] Hugo Gascon、Steffen Ullrich、Benjamin Stritter 和 Konrad Rieck。字里行间：鱼叉式网络钓鱼电子邮件的内容不可知检测。在 Michael Bailey、Thorsten Holz、Manolis Stamatogiannakis 和 Sotiris Ioannidis，编辑，攻击、入侵和防御研究，第 69–91 页，Cham，2018 年。施普林格国际出版社。

[15] John A Hartigan 和 Manchek A Wong。算法 AS 136：一种 k 均值聚类算法。皇家统计学会杂志。 C 系列（应用统计），28(1):100–108，1979。

[16] 亚历克斯·赫恩。 Unicode 技巧让黑客隐藏网络钓鱼 URL，2017 年。 https://www.theguardian.com/技术/2017/4/19/网络钓鱼-url-技巧-黑客.

[17] 埃尔南德斯。购房者在电汇欺诈交易中损失毕生积蓄，起诉富国银行，房地产经纪人和产权公司，2017 年。https://www.thedenverchannel.com/钱/消费者/购房者-失去生命-电汇欺诈交易期间的储蓄-苏威尔斯法戈房地产经纪人产权公司.

[18] Grant Ho、Asaf Cidon、Lior Gavish、Marco Schweighauser、Vern Paxson、Stefan Savage、Geoffrey M. Voelker 和 David Wagner。大规模检测和表征横向网络钓鱼。在第 26 届 USENIX 安全研讨会（USENIX 安全 19）中。 USENIX 协会，2019 年。

[19] Grant Ho、Asaf Cidon、Lior Gavish、Marco Schweighauser、Vern Paxson、Stefan Savage、Geoffrey M. Voelker 和 David Wagner。大规模检测和表征横向网络钓鱼（扩展报告）。在 arxiv，2019 年。

[20] Grant Ho、Aashish Sharma、Mobin Javed、Vern Paxson 和 David Wagner。检测企业设置中的凭据鱼叉式网络钓鱼。第 26 届 USENIX 安全研讨会（USENIX 安全 17），第 469-485 页，温哥华，不列颠哥伦比亚省，2017 年。USENIX 协会。

[21] Ling Huang、Anthony D. Joseph、Blaine Nelson、Benjamin I. P. Rubinstein 和 J. Doug Tygar。对抗性机器学习。在 AISec，2011 年。

[22] 信息安全研究所。 网络钓鱼数据 - 攻击统计数据，2016 年。http://resources.infosecinstitute.com/类别/企业/网络钓鱼/the-phishing-景观/网络钓鱼数据攻击统计/.

[23] SANS 研究所。 来自战壕：Sans 2016 年金融部门安全和风险调查，2016 年。https://www.sans.org/reading-room/白皮书/分析师/战壕-2016-调查-安全风险金融部门 37337.

[24] 娜塔莉·贾普科维奇。阶级失衡问题：意义与策略。在过程中。国际会议的。关于人工智能，2000 年。

[25] M.科罗洛夫。报告：2016 年，只有 6% 的企业使用 DMARC 电子邮件身份验证，并且只有 1.5% 的企业强制执行。https://www.csoonline.com/article/3145712/安全/.

[26] Miroslav Kubat、Robert C Holte 和 Stan Matwin。用于检测卫星雷达图像中漏油的机器学习。机器学习，30(2-3):195–215，1998。

[27] Miroslav Kubat、Stan Matwin 等。解决训练集不平衡的诅咒：单边选择。在 ICML，第 97 卷，第 179-186 页。美国纳什维尔，1997 年。

[28] M. Lan、C. L. Tan、J. Su 和 Y. Lu。用于自动文本分类的监督和传统术语加权方法。IEEE 模式分析和机器智能汇刊，31(4):721–735，2009 年 4 月。

[29] 大卫·D·刘易斯和杰森·卡特利特。用于监督学习的异构不确定性抽样。第 11 届国际机器学习会议论文集，第 148-156 页，1994 年。

[30] Charles X Ling 和 Chenghui Li。直销数据挖掘：问题与解决方案。KDD，第 98 卷，第 73–79 页，1998 年。

[31] 丹尼尔洛德。对统计垃圾邮件过滤器的好词攻击。在第二届电子邮件和反垃圾邮件会议论文集（CEAS，2005 年）中。

[32] 丹尼尔洛德和克里斯托弗米克。对抗性学习。在第十一届 ACM SIGKDD 数据挖掘知识发现国际会议记录中，KDD '05，第 641–647 页，美国纽约州纽约市，2005 年。ACM。

[33] 微软。Office 365、2019 中的反欺骗保护。https://docs.microsoft.com/en-us/office365/securitycompliance/反欺骗-保护.

[34] Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。单词和短语的分布式表示及其组合性。C. J. C. Burges、L. Bottou、M. Welling、Z. Ghahramani 和 K. Q. Weinberger，编辑，神经信息处理系统进展 26，第 3111-3119 页。柯伦联合公司，2013 年。

[35] 约阿夫·纳撒尼尔。ZeroFont 网络钓鱼：操纵字体大小以绕过 Office 365 安全性，2018 年。https://www.avanan.com/resources/ zerofont 网络钓鱼攻击.

[36] C 北方。昵称和简称查询，2017 年。https://github.com/carltonnorthern/ 昵称和小名查找.

[37] N. Papernot、P. McDaniel、S. Jha、M. Fredrikson、Z. B. Celik 和 A. Swami。深度学习在对抗环境中的局限性。2016 年 IEEE 欧洲安全与隐私研讨会 (EuroS P)，第 372–387 页，2016 年 3 月。

[38] Michael Pazzani、Christopher Merz、Patrick Murphy、Kamal Ali、Timothy Hume 和 Clifford Brunk。减少错误分类成本。在第十一届国际机器学习会议论文集中，第 217-225 页，1994 年。

[39] N.珀尔罗斯。黑客瞄准核设施、国土安全部和联邦调查局。比如说，2017 年。https://www.nytimes.com/2017/07/06/ 技术/nuclear-plant-hack-report.html.

[40] J·罗斯·昆兰。C4.5：机器学习程序。Morgan Kaufmann Publishers Inc.，美国加利福尼亚州旧金山，1993 年。

[41] J.J.罗伯茨。Facebook 和谷歌是受害者 1 亿美元的付款骗局，2017 年。http://fortune.com/ 2017/04/27/facebook-google-rimasauskas/.

[42] G. 索尔顿和 M. J. 麦吉尔。现代信息检索简介。McGraw-Hill，Inc.，美国纽约州纽约市，1986 年。

[43] Z. Song 和 N. Roussopoulos。K 最近邻搜索移动查询点。第 79-96 页，2001 年。

[44] 美国证券交易委员会。8-k 表格，2015 年。https://www.sec.gov/Archives/埃德加/数据/1511737/000157104915006288/t1501817_8k.htm.

[45] Gianluca Stringhini 和 Olivier Thonnard。那不是你：通过行为建模阻止鱼叉式网络钓鱼。在入侵和恶意软件检测及漏洞评估国际会议上，第 78-97 页。施普林格，2015 年。

[46] Andrew Trask、Phil Michalak 和 John Liu。sense2vec - 一种快速准确的神经词嵌入词义消歧方法。CoRR，abs/1511.06388，2015 年。

[47] Gary M Weiss 和 Haym Hirsh。学习预测事件序列中的罕见事件。在 KDD 中，第 359–363 页，1998 年。

[48] 科林·惠特克、布赖恩·雷纳和玛丽亚·纳齐夫。大规模自动分类钓鱼页面。在 NDSS '10，2010 年。

[49] C. Willems、T. Holz 和 F. Freiling。使用 CWSandbox 进行自动化动态恶意软件分析。IEEE 安全隐私，5(2):32–39，2007 年 3 月。

[50] Gregory L. Wittel 和 S. Felix Wu。关于攻击统计垃圾邮件过滤器。在会议记录中 *电子邮件和反垃圾邮件（CEAS），2004 年。*

[51] Gang Wu 和 Edward Y Chang。不平衡数据集学习的类边界对齐。在 ICML 2003 workshop on learning from imbalanced data sets II 中，华盛顿特区，第 49–56 页，2003 年。