HZ BOOKS
华章IT
no starch press

· 网络空间安全技术丛书 ·

基于数据科学的恶意软件分析

MALWARE DATA SCIENCE

Attack Detection and Attribution

[美] 约书亚·萨克斯 (Joshua Saxe) 希拉里·桑德斯 (Hillary Sanders) 著
何能强 严寒冰 译

机械工业出版社
China Machine Press

· 网络空间安全技术丛书 ·

基于数据科学的恶意软件分析

MALWARE
DATA
SCIENCE

Attack Detection and Attribution

[美] 约书亚 · 萨克斯 (Joshua Saxe) 希拉里 · 桑德斯 (Hillary Sanders) 著 何能强 严寒冰 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

基于数据科学的恶意软件分析 /（美）约书亚·萨克斯，（美）希拉里·桑德斯著；何能强，严寒冰译 . —北京：机械工业出版社，2020.2
（网络空间安全技术丛书）
书名原文：Malware Data Science
ISBN 978-7-111-64652-5

I. 基… II. ①约… ②希… ③何… ④严… III. 计算机网络－安全技术－研究
IV. TP393.08

中国版本图书馆 CIP 数据核字（2020）第 015083 号

本书版权登记号：图字 01-2019-5028

基于数据科学的恶意软件分析

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：梁华杰　　责任校对：李秋荣
印　　刷：北京瑞德印刷有限公司　　版　　次：2020 年 3 月第 1 版第 1 次印刷
开　　本：186mm×240mm　1/16　　印　　张：15.5
书　　号：ISBN 978-7-111-64652-5　　定　　价：79.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

華章圖書

一本打开的书，一扇开启的门，
通向科学殿堂的阶梯，托起一流人才的基石。

译 者 序

为什么要翻译 *Malware Data Science* 这本书？

在读完本书的英文版之后，我这么问自己，是什么驱使我翻译了这本书呢？总结起来有以下三个方面的原因：

首先，这本书的内容同时覆盖恶意软件和数据科学这两个方面，主题非常新颖，目前国内还没有出版过这两方面相结合的技术书籍。因此，我认为这本书有着很大的参考价值，将它翻译成中文，可以帮助国内读者系统、全面地学习如何利用数据科学的方法来分析恶意软件。

其次，这本书的两位作者都是在一线从事恶意软件数据分析工作的数据科学家，书中的内容是两位作者根据他们日常的实践工作总结出来的，其中全面介绍了恶意软件的动静态分析方法，对恶意软件之间进行关联分析的可视化方法，以及利用机器学习和深度学习方法构建实际的恶意软件检测器等内容。全书既有关于恶意软件的理论知识，也有结合丰富实际案例的实践内容，我读后收获颇丰。

最后，我所从事的就是恶意软件的治理工作，在恶意软件的分析工作方面也有多年的实践经验，对这个领域非常熟悉，在翻译原书的过程中能较为准确地把握专业内容的中文表述，希望能够通过翻译这本书为国内从事恶意软件治理工作的同行提供有益的参考。

目前，恶意软件的治理工作面临巨大的挑战，互联网上每年出现在 PC 端的恶意软件有上千万个，移动端的恶意软件也有数百万个，并且还在继续快速增长。如何对数量庞大的恶意软件进行分析，提取关键信息，确定恶意软件之间的关联关系，最终梳理出攻击团伙的线索，是恶意软件治理的从业者需要思索并试图解决的难题。

近年来，随着互联网上数据规模的急速增长，机器学习、深度学习等数据科学技术在实践中得到了迅猛发展，出现了 TensorFlow、Keras、PyTorch 等数据科学实用工具，

大大降低了在实际工作中应用数据科学技术的门槛。这使得应用数据科学技术成为大规模恶意软件分析的一条可行途径。

然而，在恶意软件分析这个专业性很强的领域应用数据科学绝非易事。如何构造适合于恶意软件的特征空间？如何建立恶意软件分析所使用的数学模型？如何对数据科学方法的作用进行评价？在应用数据科学技术对恶意软件进行分析时，这些都是需要认真考虑的优化问题，我在解决这些问题的过程中就走过一些弯路。而这些问题的答案很多都可以在这本书中找到，本书能够帮助你搭建知识框架，起到入门作用。

未来，相信在更多的专业领域都会应用数据科学技术。回想起我的模式识别课程老师曾说过，万物皆可分类，通过在事物之间寻找差异性、总结差异性、分解差异性，就可以探究事物的本质。借助数字化的处理，我们可以通过机器完成分类，通过不断的训练来优化模型，提高分类的准确性，这是非常有趣的过程，希望这本书能够激发出你的灵感。

由于知识、能力、时间有限，本书的翻译难免会有疏漏和不合理的地方，欢迎读者批评指正，也希望同行能给予宝贵的建议。如果你有任何疑问或者批评建议，可以发送邮件到 malwr_data_science@qq.com 邮箱给我们反馈。

在本书的翻译过程中，得到了机械工业出版社电工电子分社的张俊红副社长、机械工业出版社华章公司的王春华老师和梁华杰老师的大力支持，感谢他们的指导和审校工作，他们的专业性使得本书的质量得以保证。感谢我的领导严寒冰主任支持我开展基于机器学习的恶意软件检测工作，并与我一同完成了本书的翻译工作。感谢我的家人让我有精力在业余时间进行翻译工作。最后，谨以此书献给我刚满周岁的女儿文新，希望她“读万卷书，行万里路”，在成长的过程中不断丰富生命的模型，健康快乐地成长！

何能强

2019 年 10 月

序

祝贺你选择本书。你正在为自己成长为一名网络安全专家准备必需的技能。在本书中，你不仅会读到将数据科学应用于恶意软件分析的精彩介绍，还有你需要熟练掌握的必要技能和工具。

目前网络安全领域的工作岗位远远多于合格的人选，所以好消息是网络安全还是一个适合涉足的领域。坏消息是要保持最新状态所需的技能在快速变化。通常情况下，需求是发明的动力。随着对熟练网络安全专业人士的需求远远超过目前所能供应的，数据科学算法正通过提供有关网络威胁的新见解和预测来填补这一鸿沟。随着数据科学越来越多地被用于在 TB 级数据中发现威胁模式，传统的监控网络数据模型正在迅速过时。值得庆幸的是，监控报警屏幕与监控停车场的视频监控系统一样令人兴奋。

那么，数据科学究竟是什么？它如何应用于安全？正如你将在本书前言中看到的，应用于安全领域的数据科学是使用机器学习、数据挖掘和数据可视化等技术来检测发现网络威胁的艺术和科学。虽然你会发现很多来自市场驱动的关于机器学习和人工智能的夸大其词，但事实上，这些技术确实已经在当前的网络安全产品中得到了很好的应用。

例如，在当前恶意软件检测的场景中，无论是恶意软件的制作规模还是攻击者在修改恶意软件特征方面的成本，都使得基于特征的恶意软件检测方法已经过时了。相反，反病毒公司现在正在训练神经网络或其他类型的机器学习算法，使用庞大的恶意软件数据集来让这些模型和算法学习它们的特征，这样就可以在不必天天更新模型算法的情况下实现新型恶意软件变种的检测。通过结合基于特征检测和机器学习检测两方面的技术方法，就可以同时覆盖已知和未知恶意软件的检测范围。本书的两位作者 Josh 和 Hillary 都是这方面的专家，他们都有丰富的经验。

但是，恶意软件检测只是数据科学的一个用例。事实上，当我们要在网络空间中发现威胁时，老练的攻击者通常是不会遗留下可执行程序文件的。相反，他们会利用已有

的软件进行初始访问，通过漏洞利用获得的用户权限，然后使用系统工具从一台机器跳转到下一台机器。从攻击的角度来看，这种方法不会留下反病毒软件能检测到的恶意软件等工具。但是，一个好的终端日志系统或终端检测和响应（EDR）系统会捕获系统级行为日志并将日志发送到云端，分析师可以从云端尝试拼接还原入侵者的数字足迹。这种需要梳理海量数据流并不断寻找入侵模式的过程是非常适合使用数据科学来解决的问题，特别是使用统计算法的数据挖掘技术和数据可视化技术来实现。未来你可以看到越来越多的安全运营中心（SOC）采用数据挖掘和人工智能技术。这确实是剖析海量系统事件数据集来识别实际攻击的唯一方法。

网络安全正在经历技术和运营的巨大转变，而数据科学正在引领这个转变。我们很幸运，有像 Joshua Saxe 和 Hillary Sanders 这样的专家，他们不仅与我们分享他们的专业知识，并且以这样一种引人入胜的、易于理解的方式进行分享。这是你了解他们知识经验的机会，同时也是将这些知识应用到自身工作中的机会，这样你就可以领先于技术的变革和那些你有责任打败的攻击者。

Anup K. Ghosh 博士，Invincea 公司创始人

于美国华盛顿特区

前　言

如果你是在网络安全领域工作，你很可能比以往更多地使用了数据科学，即使你可能还没有意识到这一点。例如，你的反病毒产品使用数据科学算法来检测恶意软件。你的防火墙供应商可能利用数据科学算法来检测可疑的网络行为。你的安全信息和事件管理（SIEM）软件很可能使用数据科学来识别数据中的可疑趋势。不管是不是明显，整个安全产业正在越来越多地将数据科学应用于安全产品中。

高级 IT 安全专业人员正在将他们自己定制的机器学习算法集成到他们的工作流程中。例如，在最近的会议报告和新闻文章中，Target 百货公司、万事达（Mastercard）和富国银行（Wells Fargo）的安全分析师都讲述了开发定制化的数据科学技术，并将其作为安全工作的一部分。[⊖]如果你还没有赶上数据科学的潮流，那么现在就是将数据科学纳入你的安全实践来提升能力的最佳时机。

什么是数据科学

*数据科学*是一个不断增长的算法工具集合，它可以让我们通过使用统计学、数学和巧妙的统计数据可视化技术来理解和预测数据。虽然有更具体的定义，但一般来说，数据科学有三个组成部分：机器学习、数据挖掘和数据可视化。

在网络安全的场景下，机器学习算法通过学习训练数据来检测新的安全威胁。这些方法已经被证明可以检测出那些能被基于特征的传统检测技术检测出的恶意软件。数据

⊖ Target 百货公司（https://www.rsaconference.com/events/us17/agenda/sessions/6662-applied-machine-learning-defeating-modern-malicious）、万事达（https://blogs.wsj.com/cio/2017/11/15/artificial-intelligence-transforms-hacker-arsenal/）和富国银行（https://blogs.wsj.com/cio/2017/11/16/the-morning-download-first-ai-powered-cyberattacks-are-detected/）。——译者注

挖掘算法通过搜索安全数据来找出一些有趣的模式（例如，有威胁的攻击者之间的关系），这些模式可能有助于我们辨别针对自身组织的攻击活动。最后，数据可视化技术将枯燥无味的表格数据转换成图像的形式，帮助人们轻松发现有趣和可疑的趋势。我将在本书中深入讨论这三方面的技术内容，并向你展示如何使用它们。

为什么数据科学对安全性至关重要

数据科学对网络安全的未来至关重要，原因有三个：首先，安全总是与数据相关。当我们试图检测网络威胁时，我们就是在对文件、日志、网络数据包和其他结构形式的数据进行分析。传统的网络安全专家不会针对这些数据源，使用数据科学技术来进行检测。相反，他们使用文件哈希值、自定义的检测规则（如特征）和自定义的启发式方法。尽管这些技术有其优点，但是针对每一种类型的攻击，都需要人为参与的技术，这就需要太多的人为工作来跟上不断变化的网络威胁形势。近年来，数据科学技术在提升我们检测网络威胁的能力方面变得至关重要。

其次，数据科学对网络安全很重要，因为互联网上的网络攻击数量急剧增长。我们以地下黑产中的恶意软件增长情况为例。2008 年，在安全社区中所知道的恶意可执行软件大约有 100 万种。2012 年，这个数字达到了 1 亿。2018 年，安全社区已知的恶意可执行软件数量已经超过 7 亿（https://www.av-test.org/en/statistics/malware/），而且这个数字可能还会继续增长。

由于恶意软件的数量庞大，基于特征的手动检测技术已不再是能检测出所有网络攻击的合理方法。由于数据科学技术使得检测网络攻击的大部分工作自动化，并大大减少了检测这些攻击所需使用的内存，因此随着网络威胁的增长，它们在保护网络和用户方面有着巨大的潜力。

最后，无论是在安全行业的内部还是外部，数据科学是这十年的技术趋势，而且很可能在未来十年仍是如此，因此数据科学对网络安全至关重要。事实上，在任何地方都可能看到数据科学的应用，如个人语音助手（亚马逊 Echo、苹果 Siri 和谷歌 Home）、自动驾驶汽车、广告推荐系统、网页搜索引擎、医学图像分析系统和健身跟踪应用程序等。

我们可以预期数据科学驱动的系统会对法律服务、教育和其他领域产生重大影响。

由于数据科学已成为整个技术领域的关键推动因素，大学、大公司（谷歌、Facebook、微软和 IBM）和政府正在投资数十亿美元来改进数据科学工具。感谢这些投资，使得数据科学工具将更适用于解决攻击检测的难题。

将数据科学应用于恶意软件

本书侧重于将数据科学应用于恶意软件，我们将恶意软件定义成为达成恶意目的而编写的可执行程序，因为恶意软件仍然是威胁发动者在网络中获得攻击立足点的主要手段，并以此实现他们的后续目的。例如，在近年来出现的勒索软件灾难中，攻击者通常向用户发送带有恶意附件的电子邮件，使得勒索软件的可执行文件（恶意软件）被下载到用户的计算机上，然后就对用户的数据进行加密，并要求用户支付赎金来解密数据。尽管一些老练的攻击者有时为逃避恶意软件检测系统的监视而不使用恶意软件，但是在目前的网络攻击中，恶意软件仍然是攻击者主要应用的技术。

本书着眼于网络安全领域中数据科学的特定应用，而不是试图广泛地涵盖整个网络安全的数据科学，旨在更全面地展示如何将数据科学技术应用于解决重大的网络安全问题。通过了解基于数据科学的恶意软件分析，我们能够更好地将数据科学应用到其他网络安全领域，比如检测网络攻击、钓鱼邮件或可疑用户行为等。实际上，你在本书中学到的几乎所有技术都不仅适用于恶意软件检测工作，而且适用于构建一般的数据科学检测和智能系统。

本书的目标读者

本书的目标读者是那些有兴趣学习更多关于如何使用数据科学技术解决计算机安全问题的安全专业人士。如果你不了解计算机安全和数据科学，你可能会意识到自己不得不通过查找专业术语来给自己提供一些相关知识，但是你仍然可以顺利地阅读本书。如果你只对数据科学感兴趣，而对计算机安全不感兴趣，那么这本书可能不适合你。

本书的主要内容

本书的第一部分由三章组成，涵盖了理解本书后面讨论恶意软件数据科学技术所必

需的基本逆向工程概念。如果你刚接触恶意软件，请先阅读前三章。如果你是恶意软件逆向工程的老手，那么你可以跳过这些章节。

- 第 1 章　包括用于分析恶意软件文件，并发现它们如何在我们的计算机上实现恶意目的的静态分析技术。
- 第 2 章　向你简要介绍了 x86 汇编语言和如何反汇编，以及恶意软件的逆向工程。
- 第 3 章　通过讨论动态分析，对本书关于逆向工程的内容部分进行总结，其中包括在可控环境中运行恶意软件来了解其恶意行为。

第 4 章和第 5 章重点关注恶意软件的关系分析，其中包括查看恶意软件集合之间的相似性和差异性，以识别针对你组织的恶意软件攻击活动，例如由一个网络犯罪团伙控制的勒索软件活动，或针对你的组织进行的有组织有针对性的攻击活动。这些独立章节非常适合那些不仅对恶意软件检测感兴趣，而且还对提取有价值的威胁情报以追踪谁在攻击其网络感兴趣的读者。如果你对威胁情报不太感兴趣而对数据科学驱动的恶意软件检测技术更感兴趣的话，那么你可以安心地跳过这些章节。

- 第 4 章　展示了如何基于共享的属性来分析和可视化恶意软件，例如恶意软件程序都会请求的主机名。
- 第 5 章　说明如何识别和可视化恶意软件样本之间的共享代码关系，这可以帮助你识别恶意软件样本集合是否来自一个或者多个犯罪团伙。

接下来的四章涵盖了你需要了解的关于理解、应用和实现基于机器学习的恶意软件检测系统。这些章节的内容还为将机器学习应用于其他网络安全场景提供了基础。

- 第 6 章　涵盖对基本机器学习概念的容易理解的、直观的且非数学化的介绍。如果你曾学习过机器学习的相关知识，本章将便于你重温这些内容。
- 第 7 章　展示如何使用基本的统计方法来评价机器学习系统的准确性，以便选择最佳方法。
- 第 8 章　介绍了可以用来构建自身机器学习系统的开源机器学习工具，并对如何使用这些工具进行了说明。
- 第 9 章　介绍如何使用 Python 来对恶意软件威胁数据进行可视化，从而揭示攻击活动和趋势，以及如何在分析安全数据时将数据可视化集成到你的日常工作流程中。

最后三章介绍了深度学习的内容，涉及更多的数学知识，是机器学习的一个高阶领

域。深度学习是网络安全数据科学中的一个热门增长领域，这些章节为你提供了充分的入门知识。

- 第 10 章　涵盖了深度学习的基本概念。
- 第 11 章　说明了如何使用开源工具在 Python 中实现基于深度学习的恶意软件检测系统。
- 第 12 章　通过分享成为一名数据科学家的不同途径以及可以帮助你在这个领域取得成功的应有素质来对全书进行总结。
- 附录　描述了本书附带的数据和示例工具实现。

如何使用示例代码和数据

如果一本编程书没有供你使用和扩展的示例代码，那么它就是不完整的。本书每一章都附有示例代码和数据，并在附录中进行了详尽的描述。所有代码都是针对 Linux 环境中的 Python 2.7 版本编写的。要访问这些代码和数据，你可以下载一个 VirtualBox 下的 Linux 虚拟机，里面已经把程序代码、数据和所需开源工具都设置好并准备就绪，然后你就可以在自己的 VirtualBox 环境中运行。你可以从 http://www.malwaredatascience.com/ 或华章公司网站（www.hzbook.com）下载本书附带的数据，也可以从 https://www.virtualbox.org/wiki/Downloads 免费下载 VirtualBox 软件。这些代码已经在 Linux 系统中进行了测试，但是如果你希望在 Linux VirtualBox 虚拟机以外运行的话，同样的代码在 MacOS 上应该可以正常工作，在 Windows 机器上应该也可以正常运行。

如果你希望在自己的 Linux 环境中安装代码和数据，可以从 http://www.malwaredata-science.com/ 或华章公司网站（www.hzbook.com）下载。你将在可下载的归档文件中找到每个章节的目录，在每章的目录中都包含相应代码和数据的 code/ 与 data/ 目录。这些代码文件对应于每一章的代码清单或代码段，对于手头上需要处理的应用程序来说更有意义。一些代码文件与代码清单里的内容完全相同，而另一些代码文件则进行了轻微的修改，以便你可以更容易地使用参数和其他选项。代码目录附带了 pip requirements.txt 文件，其中提供了每章代码运行所依赖的开源库。要在你的机器上安装这些库，只需在每一章的 code/ 目录路径中输入 `pip -r requirements.txt` 即可。

现在你已经可以访问本书的代码和数据了，让我们开始吧。

致　谢

感谢 No Starch 出版社的 Annie Choi、Laurel Chun 和 Bill Pollock，以及我的编辑 Bart Reed。公正地说，他们应该被视为这本书的合著者。提前感谢负责印刷、运输和销售这本书的工作人员，以及负责其数字存储、转换和渲染的工程师们。感谢 Hillary Sanders 在需要的时候把她非凡的才能带到这个项目中来。感谢 Gabor Szappanos 出色而严格的技术评审。

感谢我两岁的女儿 Maya，我很高兴和她一起分享，她让这个项目大大放慢了速度。感谢 Alen Capalik、Danny Hillis、Chris Greamo、Anup Ghosh 和 Joe Levy 在过去 10 年的指导。非常感谢美国国防高级研究计划局（DARPA）和 Timothy Fraser 对本书大部分内容所基于的研究进行支持。感谢 Mandiant 公司和 Mila Parkour 提供用于本书功能性验证的 APT1 恶意软件样本。非常感谢 Python、NetworkX、matplotlib、numpy、sklearn、Keras、seaborn、pefile、icoutils、malwr.com、CuckooBox、capstone、pandas 和 sqlite 的作者们对安全及数据科学软件的免费和开源所做出的贡献。

万分感谢我的父母，Maryl Gearhart 和 Geoff Saxe，感谢他们给我介绍计算机，感谢他们无限的爱和支持。感谢 Gary Glickman 对我的爱和支持。最后，感谢我的人生伴侣 Ksenya Gurshtein 在这个过程中毫不犹豫地全力支持我。

Joshua Saxe

感谢 Josh 让我参与这本书！感谢我出色的老师 Ani Adhikari。感谢 Jacob Michelini，因为他真的想在书中找到自己的名字。

Hillary Sanders

作者简介

约书亚·萨克斯（Joshua Saxe）是专业安全企业 Sophos 的首席数据科学家，他在 Sophos 公司负责领导一个安全数据科学研究团队。他还是 Sophos 公司基于神经网络的恶意软件检测器的主要发明者，它可以保护数以千万计的 Sophos 客户防范恶意软件。在加入 Sophos 之前，他花了五年时间来管理美国国防高级研究计划局资助的美国政府安全数据研究项目。

希拉里·桑德斯（Hillary Sanders）是 Sophos 公司的高级软件工程师和数据科学家，她在为 Sophos 公司发明和产品化神经网络、机器学习和恶意软件相似性分析安全技术方面发挥了关键作用。在加入 Sophos 之前，希拉里是 Premise 数据公司的数据科学家。她经常在 Black Hat USA 和 BSides Las Vegas 等安全会议上发表演讲。她曾在加州大学伯克利分校学习统计学。

评审专家简介

Gabor Szappanos 毕业于布达佩斯罗兰大学，获得物理学学位。他的第一份工作是在计算机与自动化研究所为核电站开发诊断软件和硬件。Gabor 于 1995 年开始从事反病毒工作，并于 2001 年加入 VirusBuster，负责处理宏病毒和脚本恶意软件；在 2002 年，他成为病毒实验室的负责人。2008 年至 2016 年期间，他是反恶意软件测试标准组织（AMTSO）的董事会成员，于 2012 年加入 Sophos 并担任首席恶意软件研究员。

目　录