Journal of Security and Safety Technology 安防技术, 2015, 3(3), 24-29
Published Online September 2015 in Hans. http://www.hanspub.org/journal/jsst
http://dx.doi.org/10.12677/jsst.2015.33005

Hans汉斯

The Study of APT Security Detection Architecture and Key Technologies

Yiwen Liu, Qiong Huang, Jing Yu, Zilong Zhang

Beijing Mailbox 7223, 10, Beijing
Email: lywlyw@163.com

Received: Aug. 29th, 2015; accepted: Sep. 11th, 2015; published: Sep. 16th, 2015

Abstract

In this paper, by analyzing some drawbacks of the existing APT attack detection technology, APT security detection architecture and its key technologies have been proposed. Among this, a mining algorithm for APT time-domain association rule was given, and the security knowledge base was created using large-data analysis technology. Furthermore an APT attack detection method based on classification was proposed, which occupied significant roles in the APT security detection architecture.

Keywords

APT Attack, Large-Data Analysis, Data Mining, Knowledge Discovery, Attack Detection

APT安全检测体系架构及关键技术研究

刘怡文，黄　琼，余　静，张子龙

北京市7223信箱10分箱，北京
Email: lywlyw@163.com

收稿日期：2015年8月29日；录用日期：2015年9月11日；发布日期：2015年9月16日

摘　要

本文分析了现有APT攻击检测技术及存在的问题，提出了APT安全检测体系架构和APT检测的关键技术，

文章引用: 刘怡文, 黄琼, 余静, 张子龙. APT 安全检测体系架构及关键技术研究[J]. 安防技术, 2015, 3(3): 24-29.
http://dx.doi.org/10.12677/jsst.2015.33005

给出了APT时域关联规则挖掘算法，采用大数据分析技术构建了安全知识库；并提出了基于分类的APT攻击检测方法。

关键词

APT攻击，大数据分析，数据挖掘，知识发现，攻击检测

1. 概述

APT (Advanced Persistent Threat)是一种高级且持续的攻击。相较于一般零星的黑客攻击事件，APT攻击具有计划性、较强的针对性，并可长期潜伏。美国国家标准和技术研究院对 APT 给出了详细定义：“精通复杂技术的攻击者利用多种攻击向量(如：网络，物理和欺诈等)借助丰富资源创建机会实现自己目的”。这些目的通常包括对目标企业的信息技术架构进行篡改从而盗取数据(如将数据从内网输送到外网)，执行或阻止一项任务、程序；又或者是潜入对方架构中伺机进行偷取数据。

APT 的特点表现在 A 与 P 上，A 代表 Advanced，主要表现在攻击水平高超，即攻击行为特征难以提取、单点隐蔽性强、攻击渠道多样化、攻击空间不确定。P 代表 Persistent，主要表现在攻击过程持续时间长、攻击成功后隐藏时间长。

目前，APT 威胁愈演愈烈，其演进速度超乎想象，形态也更为多元化。2009 年至 2011 年针对政府和大型企业的 APT 攻击的数量已呈增长趋势[1]。根据《2012 年我国互联网网络安全态势综述》显示，2012 年我国境内至少有 4.1 万余台主机感染了具有 APT 特征的木马程序。2014 年也发生了多起由 APT 攻击导致的大型数据外泄事件。APT 攻击的攻击范围广、针对性强，它不仅仅局限于传统的信息网络，还会威胁工控系统、移动终端等其它信息系统，针对如能源、军工、金融、科研、大型制造、IT、政府、军事等大型组织的重要资产发动攻击。

由于 APT 攻击的对象和目的存在差异，所采用的技术和方法也存在较大的不同。目前，APT 攻击常用技术包括：水坑攻击、鱼叉式钓鱼攻击、SQL 注入攻击、0Day 漏洞利用、供应链攻击、屏幕记录、加密通信、声波通信、清除痕迹等。APT 攻击的过程一般可划分为 4 个阶段，包括搜索阶段、入侵阶段、渗透阶段、收获阶段，分别实施信息收集、单点突破、代码注入及远程控制、窃密和破坏等行动。

美国在 APT 检测和防御技术上具备一定的先进性，具备及时识别发现部分 APT 攻击的能力，并配合攻击取证了解完整的攻击过程与手段，而其他国家在这方面相对落后，发现时大都是攻击后期阶段，只能进行处置而很难分析溯源整个攻击过程与手段。

2. 现有 APT 检测技术及分析

传统的检测手段在应对 APT 攻击检测时已显得力不从心。传统的安全检测技术主要有：基于签名的检测技术(如网络入侵检测、恶意代码检测，它针对已知且长期大量传播的攻击比较有效)、主动行为防御检测技术(如杀毒厂商的桌面防御系统、杀毒软件，能实时监控进程的行为，但会影响用户使用，并存在大量的误报)。传统的检测手段主要针对已知的威胁，对于未知的漏洞、木马程序、攻击手法等，无法进行检测。

目前，APT 检测技术归纳起来主要有：虚拟执行分析检测、基于异常行为的检测、基于流量和深度内容的检测、以及多种检测方法的组合等[2] [3]。这些方法也存在如下一些不足和问题。

1) 虚拟执行分析检测：典型的代表有沙箱检测技术，它通过在虚拟环境上执行检测，基于运行行为

来判定攻击。由于 APT 的长期性，沙箱检测技术短期运行效果不显著，长期运行必然耗时、耗资源。且虚拟环境与真实环境存在差异，需要进行差异性分析。

2) 基于流量和深度内容的检测：包括全流量审计技术和内容深度分析检测技术。全流量审计是指对全流量进行协议解析和应用还原，从而检测出异常的行为；深度内容检测是指对行为内容进行深度分析，从而发现异常特征。二者均可通过建模进行异常检测。由于 APT 攻击持续的时间很长，需要对长时间内的数据流量和内容进行深入、细致的分析。全流量审计和深度内容检测目前面临的最大问题是数据处理量非常庞大，需要依靠大数据处理技术。另外，对检测内容的深度很难界定，缺乏安全检测策略和专家的经验与知识的指导。

3. APT 安全检测体系架构

现有的检测方案大都侧重于 APT 攻击的某一阶段，单一检测方法不能够对 APT 攻击生命周期(攻击各阶段)进行监测，需要研究体系化的检测架构，建立 APT 安全检测体系架构，以对抗 APT 攻击带来的挑战。

APT 安全检测体系架构，采用分层的思想，建立涵盖 APT 攻击各阶段(包括搜索阶段、入侵阶段、渗透阶段、收获阶段)、覆盖多种信息来源和协议深度的检测体系。攻击者可能会侥幸绕过某一方面的检测，但要想全面地逃避掉检测，则非常困难。

APT 安全检测体系架构如图 1 所示。

该体系架构分为 4 个层次，自底向上分述如下：

1) 数据知识层：该层包括由各种信息安全漏洞、入侵特征、攻击、情报、资源、日志、数据序列、数据流、网络和 Web 数据等海量数据构成的数据库、数据仓库和安全知识库。可以共享由重要信息系统单位、基础电信运营商、网络安全商、软件厂商、互联网企业等建立的数据源。为上层提供数据来源和知识库。

2) 资源统计层：该层着重对数据源进行元数据提取、审计和流量统计，得到特征模型、因果关系模型等。

3) 协同分析层：该层主要采用大数据分析技术进行关联分析、异常检测、全流量审计和深度内容检测。其中，APT 攻击各阶段异常检测，包含信息收集阶段的检测、入侵实施阶段的检测、木马植入及远程控制与渗透阶段的检测、隐秘通道及窃取与破坏行动阶段的检测；全流量审计和深度内容检测包含基于信息来源、目的和深层协议解析的全流量及内容审计。

4) 综合应用层：该层面向用户和具体应用，提供用户可选的全局安全检测策略、智能检测和专家系统，用专家的知识和智能进行综合检测。

综合应用层	全局安全检测策略，智能检测，专家系统
协同分析层	大数据分析，关联分析，异常检测，全流量审计，深度内容检测等
资源统计层	元数据提取，审计，流量统计
数据知识层	数据库，数据仓库，安全知识库

Figure 1. APT security detection architecture
图 1. APT 安全检测体系架构

4. APT 安全检测体系的关键技术

4.1. 采用大数据分析技术构建安全知识库

目前大数据在互联网、网络通信、网络空间安全和金融领域应用广泛，有良好的数据积累。各基础电信运营商、网络安全商、软件厂商、互联网企业、测评单位等在运营过程中已构建了各种数据库和数据源，包括漏洞库、入侵库、攻击库、情报库、资源库、数据资料库、业务库、测评用例库、测试记录库、日志库，以及原始网络数据包、数据序列、数据流等等。在多种数据库和数据源的基础上，可以构建数据仓库。

数据仓库是一个从多个数据源收集的信息存储库，是面向主题的、集成的、时变的、非易失的数据集合。数据仓库集成来自多种数据源和各个时间段的数据。它在多维空间合并数据，形成部分物化的数据立方体。

构建数据仓库要满足其关键特征：首先，要面向主题；构建的数据仓库要围绕重要的主题，如漏洞、入侵、攻击等事件，排除对于决策无用的数据，提供特定主体的简明视图。其次，要集成多个异构数据源；使用数据提取、数据清理和数据变换等技术，以及装入、刷新、元数据定义等数据仓库管理工具来加载和刷新数据，确保命名约定、编码结构、属性度量等的一致性。第三，确保时变性，数据存储从历史的角度提供信息，数据仓库中的关键结构都要隐式或显式地包含时间元素。

在数据库和数据仓库的基础上，采用大数据分析技术及数据挖掘技术构建安全知识库，主要包括发现知识，挖掘隐藏的模式和关联。知识通常用规则、模式等形式表示。安全知识库包含模式库、规则库。

这里，着重研究适用于 APT 攻击检测的关联分析技术，提出 APT 时域关联规则挖掘算法。

关联分析(Association Analysis)用来发现关联规则(Association Rule)，这些规则展示属性-值对频繁地在给定数据集中出现的条件。它是从数据库中发现知识的一类重要方法。例如，两个或多个数据项的取值一起重复出现且概率很高时，就存在某种关联，可以建立起这些数据项的关联规则。发现关联规则的目标也是发现数据集中所有的频繁模式(Frequent Pattern)，频繁模式显示了频繁地出现在给定数据集中的属性-值对之间的有趣联系。关联规则的概念最早是由 Agrawal 于 1993 年提出的，现在对它的研究已经成为数据挖掘领域最重要的研究方向。

关联规则可表示为：X→Y [a, b]，其中 X、Y 为不相交的事件集，其含义为 X 的发生将会导致 Y 的发生，X 和 Y 之间存在一种关联关系，a 为关联规则的支持度，b 为关联规则的信任度。其中，“支持度”表示该规则所代表的事例(元组)占全部事例(元组)的百分比。“信任度”表示该规则所代表事例占满足前提条件事例的百分比。关联规则发现算法就是从历史数据库中发现满足需求(a 大于最小支持度和 b 大于最小信任度)的关联规则。

Agrawal 等人提出的 Apriori 算法[4]以及 Mannila 等人提出的改进算法对关联规则的研究起到了重要的促进作用。在 APT 检测的知识发现中，许多操作型数据均与时间有关，我们需要发现事件与时间之间的关联以及基于时域的事件之间的关系等，如时域关联规则、周期关联规则。周期关联规则(时域关联规则)是指某些事件周期地(按时域)在某段时间内的发生将会导致另外一些事件的发生，由于这些事件可能只在某一段时间内发生，若采用一般的关联规则发现算法，这些规则可能因为其支持度低而被忽略。

我们采用时间分段的思想，同时，将支持度、信任度按照高、低排序，称为支持度与信任度度量序列；找出满足不同度量序列要求的时域关联规则。并提出 APT 时域关联规则挖掘算法。

APT 时域关联规则挖掘算法描述如下：

1) 时间分段：按照 APT 攻击的各阶段(包括搜索阶段、入侵阶段、渗透阶段、收获阶段)进行时间分段。参照以往 APT 攻击在搜索阶段、入侵阶段、渗透阶段、收获阶段等各阶段的数据记录情况，找出攻

击数据的分布规律。按照时间分段，整理包含流量和行为日志等审计信息的数据库、数据仓库；便于按照不同的时间分段进行搜索。

2) 基于频繁集的搜索算法：按照步骤 1 中的时间分段，采用 Agrawal 基于频繁集理论的递推方法，搜索不同时间分段的审计数据，发现频繁序列模式，把属性之间的关联和记录之间的序列模式组合成为规则，产生时域关联规则。

3) 算法改进：不设定最小支持度和最小信任度阀值；分别按照支持度与信任度的度量序列值(由低到高)，建立不同的规则库，存放相应的时域关联规则。

关联分析所涉及的技术主要包括日志采集、业务流程分析、关联分析算法引擎的建立以及持续更新的知识库。其中日志采集是关联分析的基础，算法决定关联分析的能力，而知识库的建立和持续更新才是关联分析的难点和价值所在。这需要丰富的领域经验和专家模糊知识，而不是简单地给定某个阀值就能解决的问题。

4.2. 数据挖掘技术与 APT 攻击检测技术相结合

数据挖掘技术与 APT 攻击检测技术相结合，其目的是利用得到的知识库(即频繁模式和关联规则)进行数据聚集和分类，从而实现对 APT 攻击的探测和预警。

分类是根据数据的不同特征将其划分为不同的类，这些类是事先利用训练数据建立起来的。在数据挖掘过程中，攻击检测技术被看作是一个分类问题，即，确定类标和特征集，将每个审计记录分为正常行为或攻击行为两种类型。

在 APT 攻击检测中，关联分析可用来寻找攻击者的各种攻击行为之间的相关性；利用这种相关性进行数据分类，实现对 APT 攻击的检测和预测。

将数据挖掘技术与 APT 攻击检测技术相结合，我们提出基于分类的 APT 攻击检测方法。该方法包含 2 个阶段，一是学习阶段(构建分类模型)，二是分类检测阶段(使用模型预测给定数据的类标号，检测可能的攻击)。其中最关键的一步是选择合适的系统特征。我们利用构建的知识库(包括频繁模式和关联规则)，用频繁模式作为选择特征的指标；然后用分类方法(主要包括判定树归纳、贝叶斯分类方法、基于规则的分类等)进行自动学习构建分类模型；在分类检测阶段，利用分类模型对类标号未知的审计数据进行搜索和模糊匹配，标记或预测出审计数据属于“正常”类还是“异常”类，发现可能的攻击。

目前，数据挖掘技术中影响较大的分类算法有：判定树归纳的 ID3 (称为迭代的二分器 Iterative Dichotomiser)、C4.5 (ID3 的后继)；基于后验概率贝叶斯定理的朴素贝叶斯分类和贝叶斯信念网络；神经网络的后向传播算法等[5]。如果用预测的准确率、计算速度、健壮性和可解释性来评估分类算法好坏的话，不同的分类方法各有其优劣，尚未发现有一种方法对所有数据都优于其他方法，可根据应用领域选择使用。

将数据挖掘技术与 APT 攻击检测技术相结合，其主要优点是：它能从大量的审计数据中提取有用的模式，自动生成分类模型(或称检测模型)，可用于建立 APT 攻击检测系统，使检测和防御涵盖 APT 攻击的各个阶段，达到最佳效果。

5. 应用结果

利用本文提出的 APT 检测技术，基于测评工作日志信息和历史数据库(包括测试记录库、日志库、原始网络数据包、攻击数据等)，构建了测评安全知识库；并采用基于分类的 APT 攻击检测方法，开发了辅助检测工具，可辅助实现对 APT 攻击的探测和预警。

APT 安全检测体系架构的建立和使用，还需要上层的智能分析检测、以及专家经验和知识的支持。

只有在分析检测过程中充分发挥专家的经验与人工智能，并辅以有效的工具，才能对 APT 进行有效的检测。

基金项目

“十二五”预研项目资助。

参考文献 (References)

[1] 许佳, 周丹平, 顾海东 (2014) APT 攻击及其检测技术综述. *保密科学技术*, **1**, 34-40.

[2] 周涛 (2012) 大数据与 APT 攻击检测. *信息安全与通信保密*, **7**, 22-23.

[3] 刘昕 (2014) 大数据背景下的 APT 攻击检测与防御. *电子测试*, **2**, 80-81.

[4] 周根贵 (2010) 数据仓库与数据挖掘(第 2 版). 浙江大学出版社, 杭州.

[5] Han, J.W., Kamber, M. and Pei, J., 著 (2012) 数据挖掘概念与技术(第 3 版). 范明, 孟小峰, 译, 机械工业出版社, 北京.