文章编号:1671-4229(2021)03-0020-10

基于上下文感知计算的网络攻击组织追踪方法

王　津[1]，叶晓虎[1*]，肖岩军[1]，田志宏[2]
(1. 绿盟科技集团股份有限公司，北京　100089；2. 广州大学 网络空间先进技术研究院，广东 广州　510006)

摘　要：对大部分网络监管单位和企业来说,网络安全运营很大程度上已经变为一个大数据分析和处理问题。如何从海量多模态的告警数据中快速发现高危安全事件是目前监管单位和企业的一个重要课题。文章针对这一问题提出了一种基于上下文感知计算框架的攻击组织追踪方法。首先结合上下文感知计算框架从多源威胁情报和本地沙箱告警日志中采集攻击组织相关威胁语义知识构建攻击组织知识库;然后基于大数据流式计算对实时、海量和多模态告警数据进行范式化理解和攻击链关联；结合构建的攻击组织知识库进行事件威胁语义富化和攻击组织特征关联计算,最终发现海量告警背后值得关注的攻击组织相关高危事件。经过在实际生产环境中部署系统,验证了文章提出方法的有效性。

关键词：大数据；信息安全；上下文感知计算；威胁情报；攻击组织发现
中图分类号：TP 309.2　　**文献标志码**：A

The method of cyber attack group tracking based on context-aware computing

WANG Jin[1], *YE Xiao-hu*[1*], *XIAO Yan-jun*[1], *TIAN Zhi-hong*[2]
(1. NSFOCUS Technologies Group Co., Ltd., Beijing 100089, China;
2. Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China)

Abstract: For most cyber supervision departments and groups, cyber security operations have largely become a big data analysis and processing problem. How to quickly discover high-risk cyber security incidents from the massive multi-modal alarm data is an important topic for the department concerned. This paper proposes a cyber-attack group tracking method based on context-aware computing framework. First, based on the context-aware computing framework, collect the threat context knowledge of the attack groups from multi-source threat intelligence and local sandbox alarm logs to build the attack group knowledge base; then based on the big data streaming computing, the real-time, massive, multi-modal alarm data is normalized and combined into a kill chain; combining with the attack group knowledge base to carry out event threat context enrichment and group feature similarity computation and finally discovering the high-risk incidents related to the attack group behind the massive alarms. After deploying the system in the actual production environment, the effectiveness of the method proposed in this paper is verified.

Key words: big data; cyber security; context-aware computing; treat intelligence; cyber-attack group tracking

伴随着互联网技术、云计算和大数据相关技术的快速和深入发展,网络攻击和入侵行为的频率也逐年增大。其中,针对特定目标进行有组织的攻击活动其危险性相较于其他网络攻击的危害

作者简介：王津(1991—),男,硕士研究生. E-mail:wangjin2@ nsfocus. com
* 通信作者. E-mail: yexiaohu@ nsfocus. com
引文格式：王津，叶晓虎，肖岩军，等. 基于上下文感知计算的网络攻击组织追踪方法[J]. 广州大学学报(自然科学版),2021,20(3):20-29.

性要大很多。如针对企业乃至国家层面的 APT 攻击行为就给受害目标的信息系统和数据安全带来极大的挑战[1-3]。

近年来,随着我国经济快速发展,网络经济占比越来越大。当前,应用广泛的基础软硬件安全漏洞不断被披露,具有特殊目的的黑客组织攻击活动越发频繁,各企业组织乃至国家关键信息基础设施面临的安全风险不断加大。其中,有组织的攻击破坏活动日益猖獗,大量针对我国军事机密、政府网站和关键资产、科研院所研究机密成果等敏感内容进行攻击。随着《网络安全法》的正式发布,我国已经将网络空间安全提升至国家安全的战略高度,各监管单位和企事业单位也将逐步提升网络安全监测能力和高危事件及时预警的能力。

网信办国家互联网应急中心在《2019 年中国互联网网络安全报告》中指出,在 2019 年全年,我国党政军等关键性基础设施资产和相关网站频繁遭受 DDoS 攻击,攻击组织在早期攻击阶段就可导致近 80% 的网站正常服务瘫痪;Sodinokib 以及 GandCrab、GlobeImposter 等勒索病毒成为该年度最为活跃的恶意代码家族,很多恶意代码家族甚至有超过数百种的变种出现。包括 WannaMiner 和永恒之蓝等挖矿木马的活动也非常猖獗,它们频繁利用各种安全漏洞和云网盘进行大规模扩散,其中,以 CoinMiner, Xmrig 以及 WannaMiner 这 3 个家族最为流行。此外,勒索病毒相关的攻击,特别是针对企业的勒索攻击也愈发频繁,越来越显示出高度针对性的特点。

另一类日益严重的网络入侵活动就是 APT 组织的攻击。特别是 2020 年我国疫情期间,各个 APT 组织均发起了针对性的恶意攻击活动。如来自越南的“海莲花”攻击组织,自 2012 年被发现起就开始针对包括中国、东盟等国的能源、政府及医疗行业进行攻击,2020 年 1 月,该组织针对中国网络资产进行大规模入侵活动,特别是包含疫情相关内容的网站及单位。在实际监测过程中,发现该组织经常以相关政府单位的报告文件作为诱饵进行鱼叉攻击;来自印度的“响尾蛇”攻击组织同样是 2012 年被发现,该组织长期以来专门针对巴基斯坦和中国的政府、医疗、能源及军事行业的相关网络资产进行攻击,该组织同样利用武汉疫情投递包含“申请表”字样的恶意文件进行鱼叉攻击,针对的漏洞为 CVE-2017-11882,2020 年 6 月,该组织还利用中国和印度的边境冲突问题针对我国高校、科研院所及部分政府企事业单位进行攻击,2020 年 11 月,该组织又结合第二届“一带一路”会议进行相关文件伪造的鱼叉攻击;来自韩国的 DarkHotel 攻击组织长期以来针对包括中国在内的东亚国家及东南亚国家和一小部分欧洲国家的军事、政府和能源等行业进行攻击,2020 年疫情期间,该攻击组织利用深信服 VPN 漏洞进行相关恶意软件的下发来进行入侵,至少 200 台相关 VPN 服务器失陷,大部分都属于政府机构;来自印度的“摩诃草”攻击组织自 2012 年被发现后,长期针对中国和巴基斯坦的政府、医疗及军事科研机构进行渗透攻击,2020 年疫情期间,该组织经常伪造疫情相关文件进行远控恶意代码的投递;另一个来自印度的攻击组织“蔓灵花”,自发现以来长期针对我国电力、核能及军工等行业进行攻击,2020 年第一季度,该组织同样利用疫情对我国关基资产进行大范围网络入侵攻击,主要的攻击手段为利用伪冒邮箱进行钓鱼邮件发送。

为了应对这一日益严峻的网络威胁挑战,众多企业组织、关基资产及监管部门都通过部署诸如 IPS、恶意样本沙箱和 Web 防火墙等安全防护设备来进行网络防护[4-5]。但随着部署的网络防护设备日益增多,相关的告警及日志量级也随之增长,并且由于厂商和设备的不同,日志和告警的格式也差异较大,这给相关的研判运维人员带来了极大的挑战。Gartner[6] 在 2012 年的相关报告中就指出:信息安全的问题正在转化成一个大数据的分析问题,大规模的安全数据需要被有效的关联、分析和挖掘。很多传统的安全分析技术在海量多模态数据场景下,不再能够继续满足精细化和高效的安全分析需求[7]。针对大数据场景下的这一分析问题,又可以进一步拆解成为 2 个主要的问题:①如何将接入的海量多模态告警日志数据进行理解、关联和存储;②如何构建有效的分析模型,从海量数据中快速发现真正值得关注的高威胁事件。

在海量的安全事件当中,那些由已知攻击团伙发起的事件显然是更为值得关注的。这些已知的攻击组织既包括已经在网络上被披露的 APT 攻击组织、恶意代码家族及僵尸网络等以开源或商业的威胁情报提供的内容,也包括具体业务单位内部通过长期运营发现积累的一些针对性的攻击团伙信息。威胁情报对于在大数据场景下进行网络安全事件分析起到非常重要的作用[8-12],

也有研究者结合威胁情报构建算法模型进行攻击类型判断和攻击组织画像[13-14]。结合威胁情报可以一定程度上规避传统安全防护依赖网络安全防护设备自身的规则库和信誉库所带来的局限性问题,但比起实际的网络流量,威胁情报往往存在包含关键信息过少、质量参差不齐等问题。此外,业务单位的实际流量当中包含着大量针对性的威胁特征,这些威胁特征往往比起海量的第三方威胁情报更加贴近业务单位,也更加值得关注。

因此,本文针对在海量多模态数据场景下进行攻击组织相关事件追踪这一问题,提出一种基于上下文感知计算的攻击组织追踪方法,首先构建攻击组织的威胁本体,之后基于本体设计和实现包含上下文采集、上下文推理模块的攻击组织上下文感知计算框架,从而将从多源异构情报源获取的威胁情报这一类低层上下文转化成为统一的可以相互关联的高层上下文,并结合大数据流式计算框架,对海量多模态数据进行范式化理解、关联和威胁语义的抽取,并通过快速特征关联方法发现与攻击组织相关的事件。

1 基于上下文感知计算的攻击组织追踪方法

为了有效解决海量多模态数据场景下攻击组织相关的高危安全事件的快速发现问题,本文结合基于攻击组织本体构建上下文感知计算框架,并结合大数据流式计算,进行多模态数据的范式化理解,上下文的采集、关联和特征相似度计算,最终实现从海量威胁告警中快速发现攻击组织相关的高危事件,总体的框架如图1。

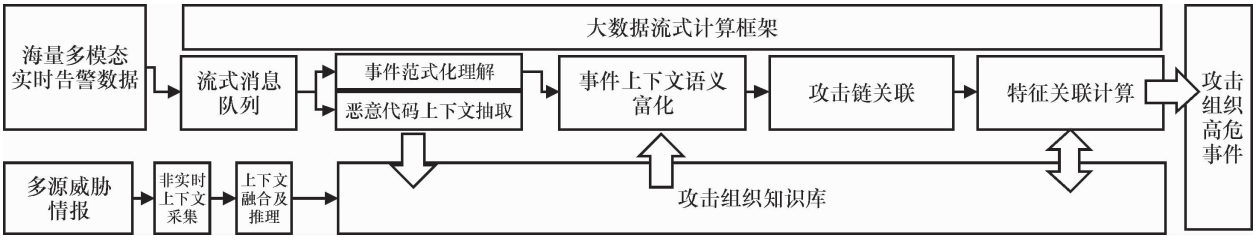

图1 基于上下文感知计算的攻击组织追踪方法总体框架

Fig. 1 The framework of cyber-attack group tracking method based on context - aware computing

如图1所示,该框架融合了包括多源威胁情报和海量多模态实时数据,一方面,针对非实时的多源威胁情报利用上下文采集、融合及推理,将攻击组织相关威胁知识存储进攻击组织知识库;另一方面,通过大数据流式计算框架结合攻击组织知识库,对实时的海量多模态数据进行理解、恶意代码语义抽取、事件上下文语义富化、攻击链关联推理和特征关联计算,发现攻击组织相关的高危事件。

1.1 基于攻击组织本体的上下文感知计算框架

构建基于攻击组织本体的上下文感知计算框架,首先需要定义以攻击组织为核心的本体结构,然后基于该本体结构设计上下文的采集模块和上下文推理模块,将非实时的多源异构威胁情报和实时的海量数据转化为攻击组织相关的关键知识,存储到攻击组织知识库中。基本框架图如图2所示。

(1)上下文采集

上下文采集模块的主要功能是从异构、复杂多样的信息源中获取上下文信息。

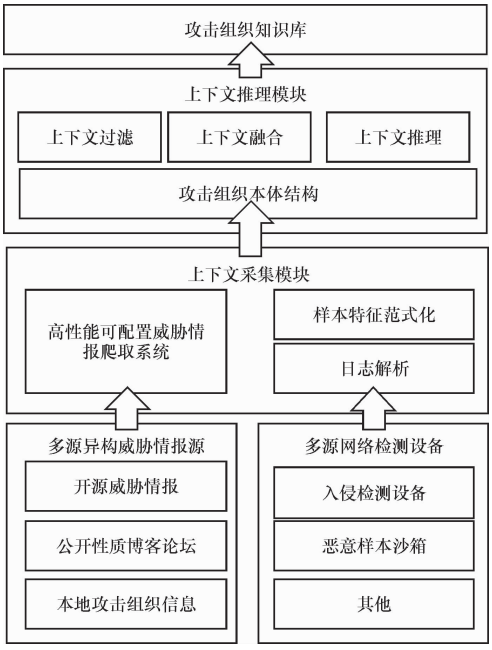

图2 基于攻击组织本体的上下文感知计算框架

Fig. 2 Context-aware computing framework based on cyber-attack group ontology

上下文采集一方面包括非实时的非结构化和半结构化网页及结构化 SITX 格式的开源威胁情报信息[15],公开性质的博客论坛以及本地积累的攻击组织威胁情报信息等;另一方面也包括网络威胁检测设备和恶意样本沙箱等实时的结构化日志告警信息。

针对非实时的多源异构威胁情报源,由于安全分析报告等情报源以非结构化和半结构化的文本为主,并且不同来源的威胁情报网页格式也差异巨大。因此,需要设计高性能可配置模板的爬虫系统,不仅可以高效地爬取网页中关键信息,同时也可以针对不同来源网页进行灵活配置,提升爬取系统的可适配性。

针对多源网络监测设备的海量实时告警,则利用大数据流式计算框架进行事件的范式化,并结合样本动作模式化进行样本静态和动态特征知识的抽取。

1)多源异构威胁情报采集

多源异构威胁情报采集主要包含 3 个关键模块:①原始情报采集模块;②情报主题区分模块;③情报自动化提取模块。详细步骤如图 3 所示。

图 3　多源威胁情报采集模块流程

Fig. 3　Multi-source threat intelligence collection process

在原始情报采集模块中,针对不同类型的情报源配置不同爬虫模板,爬虫模板可配置网页不同部分的解析规则,基于爬虫模板结合爬虫框架,通过优化爬虫框架的中间件进行指纹去重,从而提升爬虫的性能。

在情报主题区分模块中,通过对网页进行分词并结合主题词典进行词权重计算,实现快速对爬取的威胁情报进行主题划分,如 APT 组织主题、恶意代码家族主题和僵尸网络主题等。

情报自动化提取模块中,首先进一步针对爬取内容进行清洗,去除掉无效特殊字符,判断文章结尾等;之后结合正则表达式、命名实体识别技术以及关键字字典,对文本中多类威胁情报进行自动化提取和筛选。

2)多源网络检测设备实时数据采集

针对网络威胁检测设备和恶意样本检测沙箱等设备输出的实时结构化或半结构化数据,由于这一类数据通常是实时产生的、数量巨大,并且不同厂商、不同类型的设备输出数据格式也差异较大,因此,设计基于大数据计算框架的海量日志处理模块,通过流式计算,实现对海量异构数据的快速处理和范式化。

特别是针对沙箱类的样本日志告警,由于大部分包括 APT 组织和僵尸网络等攻击团伙,在进行入侵渗透攻击时,往往都需要进行恶意样本的投放,样本静态和动作特征对于进行团伙的识别具有非常大的价值,因此,本地沙箱设备捕获的样本及相关告警作为高价值的威胁上下文语义,同样需要进行采集,存入攻击组织知识库,从而支撑后续进行关联推理。具体的模块流程如图 4 所示。

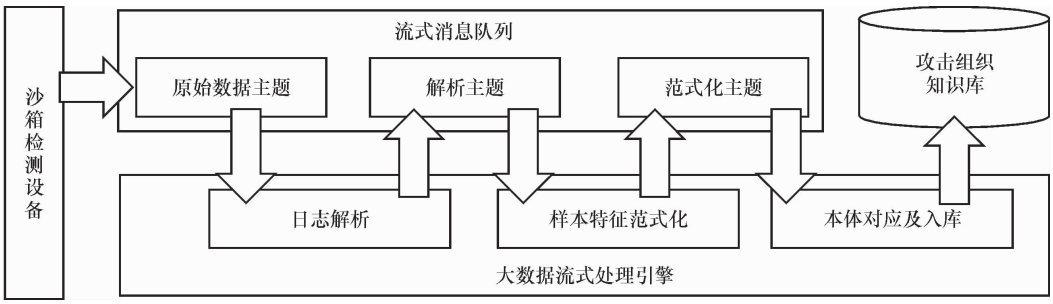

图 4　沙箱样本语义采集流程

Fig. 4　Sandbox sample context collection process

如图4所示,基于大数据流式消息队列和流式处理引擎,经过日志解析、样本特征范式化和入库几个阶段,实现将实时告警中的样本静态和动作特征采集至攻击组织知识库中。

(2)上下文推理

经过上下文采集模块获得的上下文语义信息往往有明显的异构性,其表示的方式不一致,具有单一、低层和不精确不稳定的特点。上下文推理模块的主要目标就是通过构建统一的上下文模型,将采集模块收集到的原始上下文利用过滤、推断和融合的方式,将原始的低层上下文转化成为具有统一描述格式的高层上下文,并存储在知识库当中。

1)上下文建模及攻击组织本体结构

上下文感知计算的一个基础要求是能够处理各式各样的上下文数据,这就需要建立统一的上下文表示模型。本文通过自上而下的方式定义攻击组织本体模型,并基于该本体模型进行上下文建模和表示。主要目标是构建统一的语义表示模型,一方面构建攻击组织本体结构,包括本体下各个实体概念类的定义以及实体概念类之间关系的定义;另一方面是上下文语义存储格式的统一,比如统一用key-value键值对或者是json格式等。

本文基于攻击组织所涉及的领域知识范围,定义了攻击组织本体结构,该本体包含7个实体概念:威胁主体、目标客体、攻击模式、恶意代码模式、隐患、威胁指示器和战役。各个实体概念之间的简要关系如图5所示。

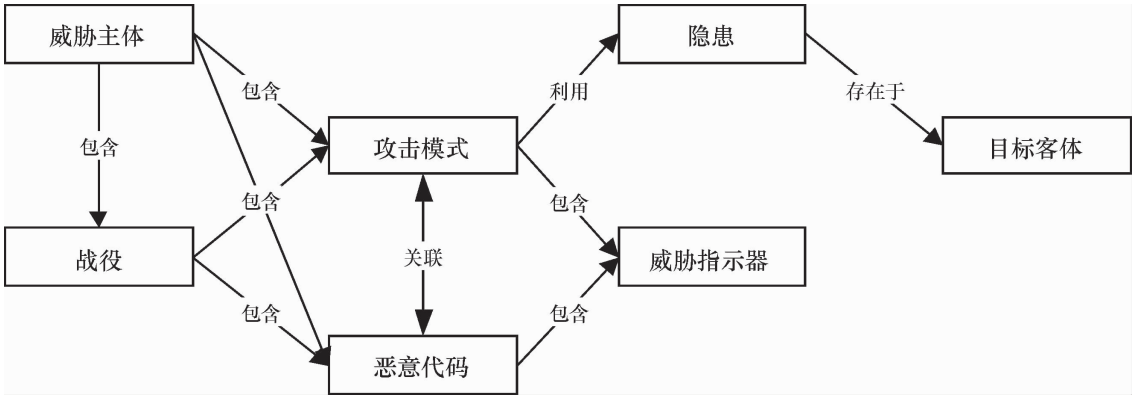

图5 攻击组织本体结构

Fig. 5 Cyber-attack group ontology

各个实体概念类的简要描述如下:

威胁主体:攻击者和攻击组织的详细信息;

战役:攻击组织实际发起的攻击行为,包含所使用的攻击模式、恶意代码等信息;

攻击模式:攻击组织进行攻击时所采用的范式化的攻击技术手段;

恶意代码:攻击组织进行攻击时所利用的恶意样本信息,包括范式化的静态信息和动作信息;

隐患:攻击模式所对应的漏洞和资产脆弱性信息;

威胁指示器:基本的网络威胁特征信息,包括恶意IP、恶意代码Hash、各种恶意代码动作参数等;

目标客体:攻击组织攻击目标资产的详细信息;

在实际使用的知识库中,各个实体概念还有很多其他的属性类型实体与之关联,这些实体之间的关系类型也更加复杂。

2)上下文过滤、融合及推理

上下文推理模块的主要功能一方面在于构建统一上下文表达模型及结构,另一方面就在于如何识别有效的上下文语义信息,实现语义的筛选过滤、相同语义的融合,以及基于利用多个上下文语义信息进行推理,从而将经过上下文采集模块获取的模糊、重复、矛盾和不精确的原始语义转化成为统一的、更加准确的高层上下文。

①上下文过滤

上下文过滤主要用于解决从原始上下文中甄选需要存储和支撑语义推理的语义内容,并且剔除一部分格式内容错误的语义。

在实际进行上下文语义过滤的过程中,首先

基于构建的上下文表示模型,即攻击组织本体结构,构建对应的实体和关系抽取规则,将规则与文本字符串进行匹配,识别命名实体。此外,基于实体抽取规则获取的内容可能会出现格式错误不合要求的情况,此时需要构建格式匹配规则,针对所有抽取的上下文语义内容,进行格式匹配,剔除掉不符合要求的语义内容。

②上下文融合

上下文抽取完成之后,由于其复杂性、多异性和模糊性的特点,导致同一实体概念可能对应多种的同义语义内容,因此,需要进行上下文语义融合,实现对同义语义内容的融合与消歧。

在实际进行上下文融合的过程中,主要采取实体链接的方法来进行。实体链接是将上下文中提到的实体与知识库中对应的实体进行链接,可以有效解决实体间的歧义性问题。

通常情况下,实体的歧义性主要表现在两个方面,首先是称之为多词同义的情况(Mention Detection,MD),即多个不同词语指代同一个实体,如美国和 USA 实际上是指同一个实体;另外就是称之为一词多义的情况(Entity Disambiguation,ED),即同一词语可能表示为多个实体,如苹果既可以指代水果也可以指代 Apple 公司。实体链接通常需要通过实体指称(Mention)的方式将具体的词语连接到知识库正确的实体上。实体链接包括两个主要的流程。

步骤 1:候选实体生成

主要采用基于实体词典的方法,通过定义各个实体的标准化词典,再通过 Trie 树(前缀树)等方法进行匹配,将自由文本中的实体指称(Entity mention)链接到知识库中对应的实体。通过字符匹配方式链接到的实体可能会包含多个,这些实体共同组成候选实体列表。

步骤 2:候选实体排序

候选实体的排序方法是目前实体链接算法研究的重点和难点,但是针对攻击组织的知识库来说,通常不会出现多个候选实体,即便出现数量也不会太多,因此,从处理效率的角度来考虑,一般直接采用抽取实体上下文信息进行相似度计算来实现候选实体的排序。如在进行 APT 组织实体链接时,可能会出现同一个 APT 组织具有多个名称的情况,此时只需要额外多抽取描述文档中提及的该组织的别名,逐个与候选实体进行匹配,选取相似度最高的一个实体进行链接即可。

③上下文推理

通过添加一系列用户定制的上下文推理规则,知识库推理机读取知识库中上下文知识与规则进行匹配,从而构建生成新的类间关系。如下面就是一条描述某 APT 组织关联上新的 C&C 地址的新的关联关系生成规则:

[ruleCC:(? group :use ? mal)(? mal :contains? action)(? action :relateTo ? ip)(? ip :type 'C&C') -> (? group:contains? ip)]

该规则描述如果某一个组织使用过某个样本,该样本具有某一个特定动作,该特定动作包含 connectTo 关连边,并且关连上 C&C 类型的 IP,那么就建立了一条 contains 的边将该 IP 和组织关联起来,图 6 是一个摩柯草攻击组织 C&C 关联边生成的示例。

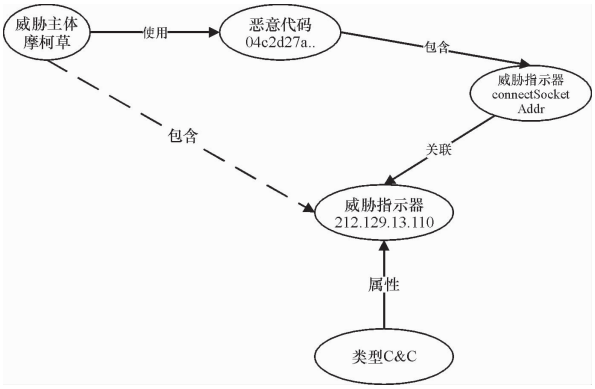

图 6　关联边推理示例

Fig. 6　Associative edge reasoning

1.2　基于事件威胁上下文的特征关联计算

(1)事件范式化理解

安全事件的范式化理解是海量实时多模态数据处理的第一步,也是后续关联推理和计算的基础。本文首先基于攻击组织本体结构定义范式化安全事件的模板,然后在大数据流式计算框架下实现流式处理引擎将海量多模态数据进行解析,最终理解成为复合安全事件模板的范式化安全事件。

1)范式化安全事件模板

从威胁主体、攻击模式和目标客体 3 个维度来定义范式化安全事件,各个维度可细化为更多的威胁特征,如表 1 所示。

表1 范式化安全事件定义模板

Table 1 Normalized event definition template

威胁特征维度	威胁特征
威胁主体	源IP
	源地理信息
	源端口
	其他信息
攻击模式	攻击模式大类
	攻击模式小类
	其他信息
目标客体	目标IP
	目标地理信息
	目标端口
	目标资产大类
	目标资产小类
	其他信息
事件信息	事件开始时间
	事件结束时间
	协议
	攻击频数
	其他信息

2)范式化安全事件模板

结合大数据流式计算框架中Spark Streaming和Kafka,基于范式化安全事件的模板,通过快速键—值映射,将海量多模态数据实时理解成为范式化的安全事件,基本流程如图7所示。

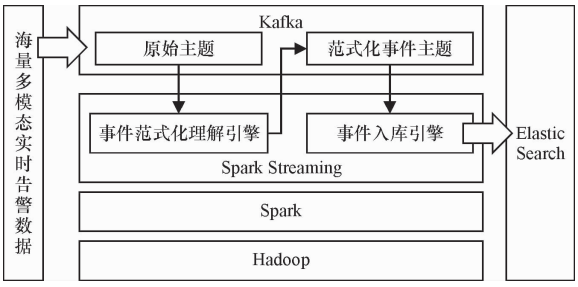

图7 范式化安全事件理解

Fig. 7 Normalized event understanding

海量多模态实时告警数据通过多种方式接入到事件范式化引擎,这些数据流被首先放入到Kafka的原始主题当中;之后基于Spark Streaming的事件范式化理解引擎,结合范式化事件模板进行映射和范式化,并将结果再次写入Kafka的范式化事件主题中,最后事件入库引擎将范式化的事件写入到Elastic Search中,供后续进行关联和推理。

(2)事件上下文语义富化

基于范式化的事件,结合攻击组织知识库分别从范式化事件3个基本维度(威胁主体、攻击模式及目标客体)进行威胁上下文语义的富化。

通过知识库的关联,可以将从多源威胁情报和本地资产情报获取的威胁主体及目标资产特征,从实时沙箱告警中获取的本地恶意代码样本静态和动作语义特征语义扩充到事件当中。事件各威胁特征维度的主要威胁语义如表2所示。

表2 事件威胁上下文语义列举

Table 2 Event threat context example

事件威胁特征维度	相关威胁语义	备注说明
威胁主体	攻击组织	哪些攻击组织使用过该主机
	IP威胁类型	C&C、恶意扫描源和僵尸主机等
	资产行业	威胁主机对应的资产所属行业
	资产名称	威胁主机对应的资产名称
	其他信息	
攻击模式	攻击链阶段	攻击模式对应的攻击链阶段
	攻击战术	攻击模式可能对应的战术
	针对漏洞	攻击模式针对的漏洞
	针对操作系统	攻击模式针对的操作系统
	其他信息	
恶意代码	样本Hash	样本MD5
	文件名	
	文件类型	样本文件类型
	家族名称	样本所属恶意代码家族名称
	样本动作A	样本最主要的3个动作及对应的参数
	A动作参数	
	样本动作B	
	B动作参数	
	样本动作C	
	C动作参数	
	其他信息	
目标客体	目标资产行业	受害主机对应的资产所属行业
	目标资产名称	受害主机对应的资产名称
	目标资产重要性	资产重要性分级
	操作系统	受害主机操作系统及版本
	漏洞	受害主机已知的漏洞
	域名	受害主机相关的域名
	其他信息	

(3)事件攻击链关联

在攻击者进行实际的入侵活动时,往往不会只利用一种攻击手段,而是在更广的时间域内利用一系列相互关联的攻击方法进行攻击,以达成

攻击目标。因此,在进行攻击行为的监测和追踪时,需要将更大时间范围内的事件进行关联,从而获得更加全面和准确的攻击行为场景。

Lockheed Martin 公司从美国军方引入的信息安全领域的攻击链(Kill chain)模型是目前最广泛运用于攻击入侵行为场景描述的模型之一[16]。攻击链模型将网络威胁入侵划分成为 7 个阶段,分别为:侦察、武器化、交货、利用、安装、命令控制和在目标活动,它将威胁安全事件按照所处的入侵阶段进行划分,是定义安全事件的一个重要属性。

本文在范式化安全事件基础上,进一步基于攻击链模型将多个事件进行关联,生成包含多个事件的攻击链。

本文中定义攻击链模型表示在同一相近时间段中,针对同一目标资产发起的一系列攻击(安全事件)。按照事件时间序列,将某一时间段内所有针对同一目标 IP 的事件基于攻击链阶段进行整合生成攻击链,具体事件插入行为链的过程如图 8 所示。

事件经过如图 8 的判断之后,会加入一条或者多条攻击链,如果无法加入至少一条攻击链,则生成一条新的攻击链,并插入该事件。

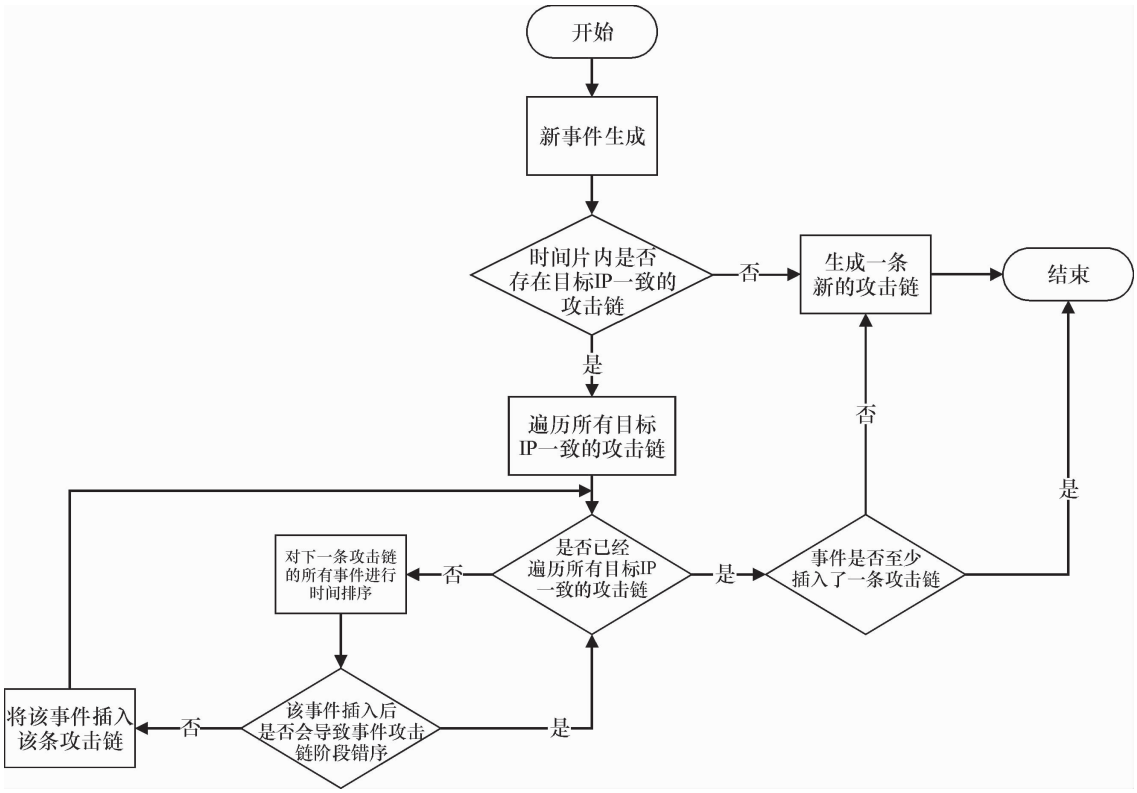

图 8　攻击链生成流程

Fig. 8　Kill-chain generation process

所有攻击链包含的事件相关的威胁主体和目标客体等语义,会同时归并生成攻击链的上下文语义。所有攻击链及相关的上下文语义共同构成复语义攻击链。

(4)攻击组织特征关联计算

复语义攻击链生成之后,需要基于攻击链相关语义进行组织的深度关联计算。关联计算的基本步骤如图 9 所示。

图 9　攻击组织语义关联计算流程

Fig. 9　Cyber-attack group context-related calculation process

如图9所示,计算复语义攻击链和攻击组织的相关度具体包含以下5个步骤:

步骤1:抽取攻击链语义特征

抽取设定周期内所有攻击链相关的事件的上下文语义,抽取的语义特征按照事件粒度进行去重操作,最后将抽取的攻击链及其语义特征以key-value格式进行暂时缓存。Key为攻击链ID,value为攻击链对应的特征列表。

步骤2:攻击链特征向量表示

基于攻击组织本体结构定义攻击组织语义特征向量模板,特征向量每一个维度对应本体一个概念实体类,将攻击链的特征列表按照特征向量模板进行表示,特征向量的每一个维度都表示为攻击链在对应概念实体类下的特征列表。

步骤3:获取语义相关攻击组织列表

基于攻击链语义特征向量进行攻击组织知识库图查询,获取所有特征相关的攻击组织列表,以及攻击组织本体特征三元组。

步骤4:攻击组织本体特征向量化

与步骤2相似,将攻击组织本体特征三元组按照定义的攻击组织语义特征向量进行重新表示,特征向量的每一个维度都表示为攻击组织在对应的概念实体类下的特征列表。

步骤5:特征相似度计算

设步骤2得到的某一条攻击链特征向量为A,A由特征列表$A_1,A_2,\cdots,A_n$组成,其中$A_1,A_2,\cdots,A_n$表示对应概念实体类的特征列表;同理,设步骤4得到某一攻击组织特征向量为B,B同样由对应概念实体的特征列表$B_1,B_2\cdots,B_n$组成,于是定义特征向量相似度算法如下:

$$S(A,B) = \sum_{i=1}^{n}\omega_i \cdot sim(A_i,B_i),\ \omega_i \in (0,1).$$

其中,ω_i表示攻击组织特征向量在某一概念实体类上的权重值,且$\sum_{i=1}^{n}\omega_i = 1$;$sim(A_i,B_i)$表示特征向量$A$和$B$在某一概念实体特征列表的相似度,按照Jaccard系数定义特征列表相似度如下:

$$sim(A,B) = \frac{|A\cap B|}{|A\cup B|},$$

经过上述步骤,可以计算出某一复语义攻击链和某一攻击组织的相关指数S,S为0到1之间的一个值,通常表示为百分比的形式。

2 系统应用情况

本文提出的基于上下文感知计算的攻击组织追踪平台已在某大型业务云上成功部署并上线运行。系统平台由3台128G内存Intel Xeon E5-2670 CPU的服务器组成分布式集群,集群支持动态节点扩容和弹性部署需求。平台接入超过3 300台不同入侵检测设备、防火墙和样本沙箱设备相关告警,日均原始告警数量超过10亿条。

系统上线运营期间,上下文采集、上下文推理、事件范式化理解、语义富化、攻击链关联和特征关联计算模块均得到实际验证,基本稳定运行;结合2021年5月某一周实际数据来看,平均每天接入告警约11亿8百万条;生成事件数目约515万;攻击链105万;其中,经过关联计算与高危攻击组织关联度较高的攻击链有236个,将海量数据压缩到了可处置的量级,极大地减轻了运维研判人员的日常运维压力。目前,该系统已支撑超过多个行业监管单位的周报和日报运维业务,反馈情况良好。

3 结 语

针对目前众多安全监管单位和企业组织在海量多模态告警数据背景下进行攻击组织活动追踪监控及相关安全事件发现的需求,本文提出了一种基于上下文感知计算的攻击组织追踪方法,创新性地设计了基于攻击组织本体的上下文感知计算框架,通过上下文采集和上下文推理模块,将非实时的多源异构威胁情报和实时的沙箱样本信息进行采集,并进行语义的过滤、融合及推理后存储至基于攻击组织本体构建的攻击组织知识库中;结合大数据流式计算框架,针对海量多模态告警数据进行事件范式化理解,基于攻击组织知识库进行事件威胁上下文语义的富化、攻击链关联和攻击组织特征关联计算,最终发现攻击组织相关的高危事件。通过在实际生产环境中进行系统的部署和运营之后,系统能够将待研判事件降低到可处置范围,有效降低研判处置人员的工作量。在接下来的工作中,可以融合现有方法的结果,通过博弈论进行安全防御策略的最优决策[17-18]。

参考文献:

[1] Chen P, Desmet L, Huygens C. A study on advanced persistent threats[C]//IFIP International Conference on Communications and Multimedia Security. Berlin, Heidelberg: Springer, 2014: 63-72.

[2] Virvilis N, Gritzalis D. The big four-what we did wrong in advanced persistent threat detection? [C]//2013 International Conference on Availability, Reliability and Security. Piscataway:IEEE, 2013: 248-254.

[3] Yang G, Tian Z, Duan W. The prevent of advanced persistent threat[J]. Journal of Chemical & Pharmaceutical Research, 2014,6(7):572-576.

[4] 杨秀云, 王玉军, 刘露. 高校云计算数据中心网络安全问题及防护措施[J]. 网络安全技术与应用, 2018,211(7): 84-85.

[5] 林玉梅. 高校校园网络安全防护方案的设计与实施[D]. 泉州:华侨大学,2015.

[6] Gartner. Information security is becoming a big data analytics problem[EB/OL]. (2012-03-23). https://www.gartner.com/doc/1960615/information-security-big-data-analytics.

[7] 陈建昌. 大数据环境下的网络安全分析[J]. 中国新通信,2013,15(17):13-16.

[8] 管磊,胡光俊,王专. 基于大数据的网络安全态势感知技术研究[J]. 信息网络安全,2016(9):45-50.

[9] 陈兴蜀,曾雪梅,王文贤,等. 基于大数据的网络安全与情报分析[J]. 工程科学与技术,2017,49(3):1-12.

[10] 张瑜,潘小明,LIU Qingzhong,等. APT 攻击与防御[J]. 清华大学学报(自然科学版),2017,57(11):1127-1133.

[11] 李超,周瑛. 大数据环境下的威胁情报分析[J]. 情报杂志,2017,36(9):24-30.

[12] 荣晓燕,宋丹娃. 基于大数据和威胁情报的网络攻击防御体系研究[J]. 信息安全研究,2019,5(5):383-387.

[13] Qamar S, Anwar Z, Rahman M A, et al. Data-driven analytics for cyber-threat intelligence and information sharing[J]. Computers & Security, 2017,67:35-58.

[14] 黄志宏,张波. 基于大数据和图社群聚类算法的攻击者画像构建[J]. 计算机应用研究,2021,38(1):232-236.

[15] Andrian J, Kamhoua C, Kiat K, et al. Cyber threat information sharing: A category-theoretic approach[C]//2017 Third International Conference on Mobile and Secure Services (MobiSecServ). Piscataway:IEEE, 2017: 1-5.

[16] Hutchins E M, Cloppert M J, Amin R M. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains[J]. Leading Issues in Information Warfare & Security Research, 2011, 1(1): 80-94.

[17] Jiang W, Fang B, Zhang H, et al. Optimal network security strengthening using attack-defense game model[C]//2009 Sixth International Conference on Information Technology: New Generations. Piscataway:IEEE, 2009: 475-480.

[18] Jiang W, Tian Z, Zhang H, et al. A stochastic game theoretic approach to attack prediction and optimal active defense strategy decision[C]//2008 IEEE International Conference on Networking, Sensing and Control. Piscataway:IEEE, 2008: 648-653.

【责任编辑: 陈　钢】