



网络安全分析中的大数据技术应用

王 帅,汪来富,金华敏,沈 军

(中国电信股份有限公司广州研究院 广州 510630)

摘 要:随着网络安全信息规模的增长,应用大数据技术进行网络安全分析成为业界研究热点。从网络安全分析的需求及传统技术的不足出发,分析了引入大数据分析的必要性,从安全数据存储、检索、分析等层面探讨了大数据技术在网络安全分析中的应用,在此基础上提出一个基于大数据的网络安全分析平台,并对典型攻击场景的关联分析方法进行了详细阐述。

关键词:大数据;安全分析;攻击检测

doi: 10.11959/j.issn.1000-0801.2015176

Big Data Application in Network Security Analysis

Wang Shuai, Wang Laifu, Jin Huamin, Shen Jun

(Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

Abstract: With the rapid growth of network security information scale, how to apply big data in network security analysis has become the network security research hotspot. From the perspective of the requirement of network security analysis and insufficient of traditional technique, the necessity of introducing big data analysis technology was analyzed, and the application of big data in network security analysis was discussed from aspects of security data storage, retrieval and analysis. On this basis, the network security analysis platform based on big data was proposed, and the correlation analysis methods of typical attack scenarios were detailed.

Key words: big data, security analysis, attack detection

1 引言

网络安全不仅涉及国家公民信息以及隐私安全,更涉及国家安全,尤其是近年来全球网络安全事故频发,美国棱镜门等事件更使得网络安全关注程度全面升级。我国对网络信息安全更为重视,2014年成立了国家安全委员会,网络及信息安全正式提升为国家战略。但与此同时,我国网络信息安全形势仍十分严峻,安全攻击事件仍呈高发态势,分布式拒绝服务(distributed denial of service,DDoS)攻

击、木马僵尸网络、移动恶意程序、用户敏感信息窃取等各类网络攻击事件数量均位居世界前列。DDoS、木马僵尸网络、移动恶意程序、APT(advanced persistent threat)等也成为电信运营商的顽疾,不仅阻塞网络带宽,降低服务质量,而且严重影响企业声誉。在安全威胁具有更强的杀伤力与逃避能力的形势下,仅靠防范措施已无法应对安全威胁,精确检测分析和早期预警成为新一代安全能力的关键。

传统基于数据的安全分析技术已有广泛应用,如恶意

收稿日期:2015-06-04;修回日期:2015-07-09

论文引用格式:王帅,汪来富,金华敏等. 网络安全分析中的大数据技术应用. 电信科学,2015176

Wang S, Wang L F, Jin H M, et al. Big data application in network security analysis. Telecommunications Science, 2015176



代码检测、入侵检测、安全审计等,但随着网络宽带化、应用丰富化等趋势,各种类型的安全数据越来越多,传统分析方法存在不足,同时新兴威胁与内控/合规的深入也对安全检测分析提出更高要求。大数据具有 volume、variety、velocity、value 的“4V”特性,可实现大容量、低成本、高效率的网络安全分析能力,能够满足海量安全信息的处理和分析需求,如何将大数据技术应用于信息安全领域已成为国内外研究热点。

本文首先分析了传统网络安全分析技术的现状和不足以及引入大数据进行网络安全分析的优势,阐述在网络安全分析中如何引入大数据技术,提出一个基于大数据的安全分析平台框架,并深入探讨了几个常见攻击场景的大数据分析方法。

2 网络安全分析引入大数据技术的必要性

随着网络架构日趋复杂,网络安全分析的数据量呈指数级增长,从 TB 级向 PB 级迈进且数据来源丰富,内容更为细化,数据分析的维度更为广泛;同时随着设备性能的增长,数据源发送速率更快,对安全信息采集的速度要求越来越高,0 day 等漏洞日益增多,影响也更加广泛,需要能够更快更及时地检测出安全隐患及安全攻击行为。此外,有目标、有组织、长期潜伏渗透的多阶段组合式 APT 攻击更为猖獗,需要分析种类更多的安全信息,并采用多种手段进行融合检测。

在传统技术架构下,采用结构化数据库进行数据存储,由于数据存储成本高,往往将原始数据进行标准化处理后再进行存储,导致部分信息丢失,长期历史数据难以保存,追踪溯源困难;同时对不完整,包含嘈杂的大型、非结构化数据集上执行分析和复杂查询的效率很低,实效性和准确性难以保证;此外,难以对来源于不同系统的异构数据进行融合关联分析,难以形成全局的可视化安全态势,安全运营效率不高。因此传统安全技术难以满足网络安全分析的新要求。

Gartner 在 2013 年的资料显示,大数据技术作为未来信息架构发展的十大趋势之首,近年在多个领域的应用发展迅速。其具有数据量大、种类繁多、速度快、价值密度低等特点,应用于网络安全分析领域可实现容量大、成本低、效率高的网络安全分析能力。

- 容量大:大数据支持海量异构数据的存储与计算,使海量原始安全信息的存储和分析成为可能。

- 成本低:分布式数据库相对传统数据库价格大大降低,且易于在低廉硬件上水平扩展,降低安全投入。
- 速度快:异构数据的存储与查询速度更快,从而使安全信息采集、检测响应速度更快。
- 精度高:数据挖掘能力提升,异构及海量数据存储的支持奠定多维多阶段关联分析的基础,基于更长时间的数据及数据间的直接关联关系,提升分析的广度和深度。

3 大数据技术在网络安全分析中的应用

目前网络安全分析主要基于日志与流量这两大类数据,同时辅以资产、漏洞、配置、访问、用户行为、应用行为、业务行为、外部情报等信息进行关联分析。其中主要基于日志的安全分析应用包括安全审计、主机入侵检测、网络安全风险分析等;主要基于流量的安全分析应用包括恶意代码检测、僵尸蠕检测、异常流量检测、Web 安全分析、网络入侵检测等。引入大数据分析技术,就是在这些安全应用的基础上,将分散日志和数据整合起来,通过高效采集、存储、检索与分析,提高安全检测与处理的效率,减少分析与响应时间;同时通过大量安全信息的关联分析、多阶段组合关联分析、攻击场景关联分析等技术,找出安全事件之间的联系和异常,有效发现 0 day、APT 等攻击和数据泄露,提升安全防御的主动性。

在安全数据采集方面,为支持海量数据的高速采集,对于日志信息可采用 Chukwa、Flume、Scribe 等工具,通过分布式采集实现每秒数百兆的日志数据采集与传输;对于流量信息可采用传统流量镜像方式,实现全流量数据采集。

在安全数据的存储方面,由于安全数据种类繁多,应用方式也不尽相同,为了满足不同分析需求的数据存储,提高分析与查询效率,对不同类型和用途的数据应采用不同的存储方式,主要包括以下内容。

- 对用于查询检索的原始日志信息、流量历史数据等信息,可采用 HBase、GBase 等列式存储方式,利用其快速索引的特点,提供对安全数据查询的快速响应。
- 对用于进行挖掘分析的标准化处理后的日志信息、流量信息等,可基于 Hadoop 分布式计算架构,将分析的数据存放于各个计算节点,通过定制化的 MapReduce 规则类与 Hive 分析脚本,对安全数据进行挖掘分析,形成分析告警与统计报表,分析的结果存放到列式存储中。

- 对用于实时分析的日志、流量等信息,可采用 Storm、Spark 等流式计算架构,将分析的数据存放到各个分析节点,通过对实施事件流的关联分析,生成安全告警与统计数据,分析结果存放到列式存储中。

在安全数据检索方面,可采用基于 MapReduce 的查询检索框架,将数据查询请求分布到不同的计算节点并行处理,并最终装配每个节点的计算结果,通过分布式并行计算提升安全数据检索效率,可使对海量原始安全数据的条件化查询速度提升到秒级。

在安全数据分析方面,对于数据实时分析,可在 Storm 或 Spark 等流式计算框架基础上,结合 CEP (complex event processing, 复杂事件处理) 等技术以及特定的关联分析方法,通过数据内存实时分析、海量安全信息实时监控与关联分析,及时发现安全异常行为;对于统计分析、历史数据分析等非实时分析,可基于 Hadoop 架构,通过 HDFS 分布式数据存储及 MapReduce 的分布式计算与调度,结合数据抽取、数据聚合、数据挖掘等技术,实现风险统计、态势分析、攻击溯源等离线分析。

在多源数据及多阶段组合关联分析方面,由于大数据存储分析速度的提升,使得在短时间内对多源异构数据的挖掘分析、大规模系统安全隐患关联挖掘、攻击不同阶段间行为特征关联等成为可能。如对僵尸网络的检测分析,

不仅可结合流量与 DNS 访问特征,还可进一步扩充分析的数据源,结合全分组数据、攻击溯源数据、蜜罐数据、外部情报等信息进行深度的关联分析;又如检测到某个主机存在漏洞或遭受攻击时,可进一步关系统中其他主机是否存在相同漏洞或攻击,是否存在跳板主机进行内网渗透等行为,进而及时发现安全隐患,提早采取防范措施。

4 基于大数据的网络安全分析平台

从上述分析可以看出,通过底层大数据采集、存储、分析、检索等技术的引入,可以切实提高安全分析的效率,但如何从海量安全数据中精确挖掘出安全问题和隐患,重点还在于针对不同攻击场景设计针对性的关联分析方法,并基于大数据存储与分析的优势,从更多样、更丰富的数据源中进行更深度的关联分析。本节提出了一个基于大数据技术进行存储分析的网络安全分析平台,阐述了该平台采用的关键技术,并基于该架构提出了在僵尸网络检测与 DDoS 攻击溯源这两种典型场景下的安全数据挖掘分析方法。

4.1 基于大数据的网络安全分析平台架构

本文提出的基于大数据的网络安全分析平台整体架构如图 1 所示。

该平台自下而上分为数据采集层、大数据存储层、数据挖掘分析层、数据呈现层,其中数据采集层对基于流、事件、用户身份及威胁情报等多源异构信息进行分布式采

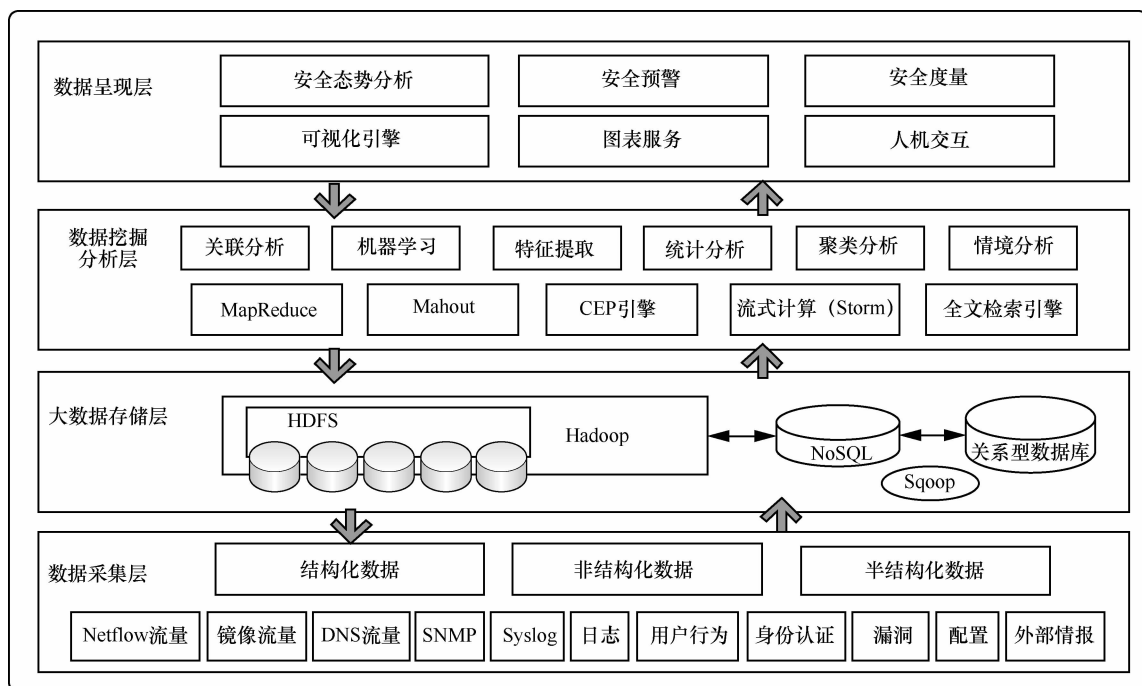


图 1 基于大数据的网络安全分析平台



集,包括 Netflow 流量、镜像流量、DNS 流量、设备 SNMP 信息、安全设备/网络设备 Syslog、Web 应用/数据库/业务日志、漏洞信息、配置信息、用户行为信息、外部情报信息等;大数据存储层利用分布式文件系统实现对海量安全基础信息的长期全量存储,提供结构化、半结构化和非结构化数据的统一存储能力,并通过均衡算法实现数据在分布式文件系统上的均衡分布,提高后续的并发检索速度;数据挖掘分析层基于 MapReduce、Mahout、CEP 等引擎与 Storm 流式计算框架,实现分布式数据关联分析、特征提取、统计分析、情境分析等安全事件挖掘能力,及时发现异常及非法网络行为并进行攻击溯源等,同时提供实时检索与溯源能力,对所有存储的海量信息进行索引,提供类似互联网搜索引擎的根据任意关键字定位到所有相关信息的能力;数据呈现层结合大数据分析技术和数据可视化技术,可以通过不同维度展现整体或局部安全态势。

4.2 平台实现关键技术

(1)数据采集技术

本平台包括对原始及初步分析后的日志与流量信息的离线统计分析与实时在线分析,离线统计分析可借助于 Hive 来实现,实时在线分析可使用实时处理系统 Storm。为了统一离线和实时计算信息的采集,本平台采用了 Flume 与 Kafka、Storm 结合的方式。

其中,采用 Flume 实现分布式、可靠和高可用的海量日志采集、聚合和传输,通过定制数据发送方收集不同数据源的数据,对数据进行简单处理后发送到各种定制的数据接收方。Flume 逻辑上分为 3 层架构:agent、collector、storage。agent 用于采集数据,是 Flume 中产生数据流的地方,同时将产生的数据流传输到 collector;collector 的作用是将多个 agent 的数据汇总后,加载到 storage 中;storage 是存储系统,可以是一个普通文件,也可以是 HDFS、HBase 等分布式文件系统。

为了处理活跃的流式数据,在数据采集与流式数据处理之间采用 Kafka 作为缓存。一种作为一种高吞吐量的分布式发布订阅消息系统,Kafka 包括多个生产者(producer)、代理(broker)和消费者(consumer),作为一个逻辑整体对外提供服务。其中 producer 负责向一个主题发布消息,consumer 接收所订阅主题的消息的广播。数据从 producer 推送到 broker,而 consumer 从 broker 上拉取数据。为了解决分布式应用中的数据管理问题,Kafka 中采用了 Zookeeper 框架来管理集群配置,进行负载均衡。

为实现对实时数据的存储分析,在数据采集与流式数据缓存之后采用了 Storm 流式计算框架,实现分布式实时消息处理。Storm 集群主要由一个主节点(运行 Nimbus)和一群工作节点(运行 Supervisor)组成,通过 Zookeeper 进行协调。Storm 提供了一系列的基本元素用于进行计算:Topology、Stream、Spout、Bolt 等,一个实时应用的计算任务被打包作为 Topology 发布,Topology 由不同的 Spouts 和 Bolts 通过数据流(stream)连接起来,其中 Spout 是 Storm 中的消息源,一般是从外部数据源(如 message queue、RDBMS、NoSQL、realtime log)不间断地读取数据并发送给 Topology 消息;Bolt 是 Storm 中的消息处理者,可以执行过滤、聚合、查询数据库等操作。

本平台中,Flume 采集数据后将离线分析的数据存储到 HDFS 中,对于实时数据流,将 Flume 作为 Kafka 消息的 producer 输入 Kafka 中,然后通过订阅的方式,使用 Storm 的 Topology 作为 Kafka 消息的 consumer,从而通过 Storm 实现对数据的实时分析处理。

(2)数据存储技术

采集后的海量数据主要存储于 HDFS 中,HDFS 是具有高容错性与高吞吐量数据访问的分布式文件系统,通过元数据节点管理文件系统的命名空间,数据节点用于存储数据文件,其默认的最基本的存储单位是 64 MB 的数据块。数据文件越小,则元数据节点存储的元数据信息记录块的位置越多,同时如果访问的文件越多,数据节点之间的跳转就越多,将严重影响性能。因此为了保证数据处理效率,本平台采用了 HDFS 的默认数据块存储单位,对采集上来的小数据文件进行归并处理,满足每个文件 64 MB 的存储大小要求。

HBase 是运行在 HDFS 上的分布式非结构化数据库,其采用列式存储的方式,实现高效大数据量的存储管理。本平台采用 HBase 存储大量非结构化数据,部分分析处理后的结果数据存储的关系型数据库中,关系型数据库与非关系型数据库间采用 Sqoop 进行数据的相互导入导出。

(3)数据分析技术

MapReduce 是一种用于大规模数据集的并行运算模型,实现数据划分、计算任务调度与分布式计算能力,其角色包括 JobTracker 与 TaskTracker,JobTracker 进行作业初始化、分配与协调,在物理位置上一般与 HDFS 的元数据节点合设;TaskTracker 负责在分配的数据片段上执行计算任

务,在物理位置上一般与 HDFS 的数据节点合设。本平台中采用 MapReduce 进行分布式数据采集、数据存储、数据查询与分析等任务的分布式调度与计算,并基于 HBase 的 Native Java API 对 HBase 表数据进行并行批处理。

本平台采用了 Hive 进行数据统计分析,使用类 SQL 的 HiveQL 语言实现 HDFS、HBase 中非结构化数据的快速查询,平台通过对 Hive 进行 API 封装,通过插件的定制开发实现对各类数据的统计分析。

为了实现数据挖掘分析,本平台采用了 Mahout 实现基于 Hadoop 的机器学习和并行数据挖掘,提供了分布式协同过滤、分类、聚类、关联规则挖掘、回归等数据挖掘算法。

为了实现对事件流的关联分析,本平台采用 CPE 技术,将系统数据看作不同类型的事件,通过分析事件间的关系,建立不同的事件关系序列库,利用过滤、关联、聚合等技术,由简单事件产生高级事件或流程,从而从海量安全信息中挖掘出安全事件。

4.3 典型实施场景

4.3.1 被入侵主机与控制源检测

基于该平台,以 DDoS 攻击历史信息、Netflow 原始数据、路由器配置数据、僵尸蠕虫检测事件、疑似僵尸蠕虫事件、僵尸蠕虫特征库、DNS 访问日志、DPI 检测到的 DDoS 攻击事件、蜜罐系统检测到的木马事件、蠕虫事件、恶意 URL 事件、外部渠道获取的僵尸蠕虫源 IP 地址、垃圾邮件源 IP 地址等信息为输入源,通过各类信息综合关联分析,并在平台中维护更新被入侵主机 IP 库、CC 主机 IP 库、恶意 URL 库、代码样本库等,实现基于大数据分析的被入侵主机和控制源检测。

具体方法如下。

(1)在 DDoS 攻击历史信息中从发动攻击的 IP 地址列表中筛选出真实地址攻击的 IP 地址,列为疑似被入侵主机。

(2)根据疑似被入侵主机 IP 地址,在 DNS 日志中查询其日志记录,并与恶意 URL 库比对,如果存在匹配记录,可基本确认其为被控制主机。

(3)在 DPI 信息中查找其与 CC 主机的通信记录或访问恶意 URL 记录,如果存在匹配记录,可基本确认为被控制主机。

(4)CC 主机定位分析:在单次攻击时间段期间,从发动攻击的所有源 IP 地址的 Netflow 信息中,关联查询出与参与此次攻击的大多数攻击源 IP 地址有交互网络流量的

IP 地址(不包括被攻击目的地址),作为疑似控制主机做进一步分析。

(5)根据疑似控制主机的 Netflow 信息,查询与控制主机有通信记录的主机,即被控制主机。

(6)对僵尸蠕虫系统、蜜罐系统、DPI 等检测到的或 CERT (computer emergency response team)等外部渠道获取的僵尸蠕虫事件信息,提取肉鸡 IP 地址列表,进一步在 DNS 访问记录中查询对应 IP 地址是否有查询恶意 URL 的记录,并在 DDoS 攻击历史信息中查询是否有攻击行为,从而对分散系统的检测结果进行综合验证,提升检测结果的准确性。

4.3.2 DDoS 攻击路径准实时监测

基于大数据网络安全分析平台,以 DDoS 攻击历史数据、Netflow 历史流量数据、网络拓扑信息、路由器配置信息等作为输入源,实现攻击准实时监测和全景溯源,并以可视化方式直观展现攻击流量穿行线路,重演攻击路径。

具体方法如下。

(1)新建空的活节点队列 Q 和空的攻击路径链表 L 。

(2)从攻击告警中提取攻击源 IP 地址/段、攻击目的 IP 地址/段、攻击时间等信息,并根据提取信息及存储 Netflow 数据关联查询攻击流量信息,包括攻击所流经路由器、路由器端口及攻击流量大小。

(3)确定攻击源路由器及端口以及攻击目的路由器及端口信息:如路由器用户侧端口输出流量含有攻击目的地址,则为攻击目的路由器及端口;如路由器用户侧端口输入流量含有攻击源地址,则为攻击源路由器及端口;并将攻击源子网与攻击源路由器连接关系以及攻击目的子网与攻击目的路由器连接关系放入攻击路径链表 L 。

(4)从攻击目的路由器 R 开始访问,此时节点 R 是已被访问的节点;以节点 R 作为扩展节点,根据网络拓扑关系,搜索所有与 R 相邻的节点 W :

- 若 W 未被访问且在 W 与 R 相连端口输出流量中查询到攻击源到攻击目的的流量,则将 W 放入活节点队列 Q 的对尾,将 W 打上已被访问的标记;并将 R 与 W 连接关系放入攻击路径链表 L ;
- 若 W 未被访问且在 W 与 R 相连端口输入流量中未查询到攻击源到攻击目的的流量,则将 W 打上已被访问的标记。

(5)访问与 R 相邻的所有节点后,再取出活节点队列 Q 的第一元素 U 作为扩展节点,如同对扩展节点 R 的处理算法一样,处理扩展节点 U ;循环处理直到活节点队列空



时为止。

(6)攻击流量穿行路径可根据攻击路径链表 L 得出。

5 结束语

Gartner 在 2012 年即指出,信息安全正在成为一个大数据分析问题。随着信息安全数据规模日益增长、类型日益繁多,现有安全分析技术难以满足精细化的高效安全分析需求。大数据高速、海量、灵活等特点易于实现大容量、低成本、精确、快速的网络安全分析能力,将大数据技术应用于网络安全分析领域是当前业界研究的一个热点。本文从当前网络攻击形势及防御需求出发,分析了传统网络安全分析技术的不足及应用大数据技术的优势,探讨了大数据在网络安全数据存储、检索、分析等方面的应用技术及方案,并在此基础上提出了一种基于大数据的网络安全分析平台以及僵尸网络检测与 DDoS 攻击溯源这两种典型场景的大数据分析方法,通过应用大数据技术,有效提升网络安全分析效率。

参考文献

- 1 孙大为,张广艳,郑纬民.大数据流式计算:关键技术及系统实例.软件学报,2014,25(4):839~862
Sun D W, Zhang G Y, Zheng W M. Big data stream computing: Technologies and instances. Journal of Software, 2014, 25 (4): 839~862
- 2 White T. Hadoop 权威指南.周敏奇,王晓玲,金澈清等译.北京:清华大学出版社,2011
White T. Hadoop: The Definitive Guide. Translated by Zhou M Q, Wang X L, Jin C Q, et al. Beijing: Tsinghua University Press, 2011
- 3 杨巨龙.大数据技术全解——基础、设计、开发与实践.北京:电子工业出版社,2014
Yang J L. Big Data Technical Total Solution——Foundation, Design, Development and Practice. Beijing: Electronics Industry Press, 2014
- 4 唐宏,罗志强,沈军.僵尸网络 DDoS 攻击主动防御技术研究与应用.电信技术,2014(11)
Tang H, Luo Z Q, Shen J. Research and application of the active defense technology of DDoS attack using botnet. Telecommunication Technology, 2014(11)
- 5 诸葛建伟,韩心慧,周勇林等.僵尸网络研究.软件学报,2008(3)
Zhuge J W, Han X H, Zhou Y L, et al. Botnet research. Journal of Software, 2008(3)
- 6 程学旗,靳小龙,王元卓等.大数据系统和分析技术综述.软件学报,2014(9)

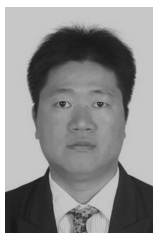
Cheng X Q, Jin X L, Wang Y Z, et al. Key laboratory of network data science and technology. Journal of Software, 2014(9)

- 7 Agrawal D, Bernstein P, Bertino E, et al. Challenges and opportunities with big data. <http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf>, 2015
- 8 韦勇,连一峰,冯登国.基于信息融合的网络安全态势评估模型.计算机研究与发展,2009(3)
Wei Y, Lian Y F, Feng D G. A network security situation awareness model based on information fusion. Journal of Computer Research and Development, 2009(3)
- 9 陈明奇,姜禾,张娟等.大数据时代的美国信息网络安全新战略分析.信息网络安全,2012(8)
Chen M Q, Jiang H, Zhang J, et al. Analysis of the U.S. information network security strategy in the era of big data. Netinfo Security, 2012(8)

作者简介



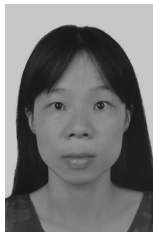
王帅,女,中国电信股份有限公司广州研究院高级工程师,主要研究方向为大数据安全、云计算安全、网络与信息安全体系及攻防技术。



汪来富,男,中国电信股份有限公司广州研究院高级工程师,主要研究方向为大数据安全、云计算安全、网络安全。



金华敏,男,中国电信股份有限公司广州研究院高级工程师,主要研究方向为 IP 网络、云计算、大数据安全、网络安全。



沈军,女,中国电信股份有限公司广州研究院高级工程师,主要研究方向为云计算大数据安全及移动互联网安全。