软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1188−1198 [doi: 10.13328/j.cnki.jos.004960]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

多视图合作的网络流量时序数据可视分析*

赵 颖[1], 王 权[1], 黄叶子[2], 吴 青[3], 张 胜[1]

[1](中南大学 信息科学与工程学院,湖南 长沙 410083)

[2](中南大学 软件学院,湖南 长沙 410075)

[3](中南大学 信息与网络中心,湖南 长沙 410083)

通讯作者: 吴青, E-mail: wuqing@csu.edu.cn

摘 要: 网络安全可视化作为一个交叉应用研究领域,为传统网络安全数据分析方法注入了新的活力.但已有研究过于注重网络安全数据的可视表达,而忽视了对分析流程的支持.抽象了网络安全分析人员用网络流量时序数据检测网络异常的过程,提出了一个自顶向下的网络流量时序分析流程模型.以该模型为指导,设计并实现了一个多视图合作的网络流量时序数据可视分析原型系统.在分析端口扫描和 DDoS 攻击等常见网络异常的案例中,该系统中的 4 个协同交互、简单易用的可视视图,可以较好地支撑分析人员由整体到个体、由点到面以及由历史到未来的网络流量时序数据分析过程.

关键词: 网络安全可视化;可视分析;网络流量;时序数据;异常检测

中图法分类号: TP393

中文引用格式: 赵颖,王权,黄叶子,吴青,张胜.多视图合作的网络流量时序数据可视分析.软件学报,2016,27(5):1188−1198. http://www.jos.org.cn/1000-9825/4960.htm

英文引用格式: Zhao Y, Wang Q, Huang YZ, Wu Q, Zhang S. Collaborative visual analytics for network traffic time-series data with multiple views. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1188−1198 (in Chinese). http://www.jos.org.cn/1000-9825/4960.htm

Collaborative Visual Analytics for Network Traffic Time-Series Data with Multiple Views

ZHAO Ying[1], WANG Quan[1], HUANG Ye-Zi[2], WU Qing[3], ZHANG Sheng[1]

[1](School of Information Science and Engineering, Central South University, Changsha 410083, China)

[2](School of Software, Central South University, Changsha 410075, China)

[3](Information and Network Center, Central South University, Changsha 410083, China)

Abstract: Cyber security visualization is a multi-discipline research field. Visualization techniques have injected new vitality into traditional analysis methods for cyber security. However, most existing studies focus on the visual expression and overlook the visual support for the data analysis process. This paper presents a top-down model for anomaly detection on network traffic time-series data drawing from the experience of cyber security analysts. A prototype system is designed based on this model, and it includes four collaborative views with direct and rich interactions. A number of experiments, including port scanning and DDoS attacking, are carried out to demonstrate that this system can support network traffic time-series analysis on overview to detail, point to area and past to future process flows.

Key words: cyber security visualization; visual analytics; network traffic; time series data; anomaly detection

随着网络通信技术的不断进步,网络应用进入了飞速发展的时代.与此同时,软硬件故障、蓄意攻击等各类

* 基金项目: 国家自然科学基金(61103108, 61402540); 湖南省科技支撑计划(2014GK3049)

Foundation item: National Natural Science Foundation of China (61103108, 61402540); Hu'nan Provincial Science and Technology Foundation (2014GK3049)

收稿时间: 2015-07-31; 修改时间: 2015-09-19; 采用时间: 2015-11-10

突发事件对网络安全的冲击也愈演愈烈.在网络运行过程中,广泛应用的网络安全产品,如流量监控系统、防火墙、入侵检测系统等,会产生大量的监控数据,这些数据是网络安全分析人员掌握网络状态和识别网络入侵的主要信息来源.许多自动化和半自动化方法被用来处理网络安全数据,但它们犹如一个黑盒子,让网络安全分析人员的经验和专业知识失去了施展空间;同时,大量的漏报和误报也让这些方法的实用价值大打折扣.

为了突出人在判断和决策上的主导地位,减轻网络安全分析人员的认知负担,有些研究者[1,2]主张将网络安全数据以可交互的图形图像的方式表现出来,借助人的视觉通道和认知能力,帮助网络安全分析人员感知和理解网络安全问题,并逐渐形成了网络安全可视化这一新兴的交叉研究领域.纵观已有的网络安全可视化研究,往往过于注重网络安全数据的图形化表示方法,这些研究成果虽然能够有效地检测某些特定攻击类型,也可以辅助分析人员快速地完成异常主机定位和攻击特征识别等特定分析环节,但很少从分析人员的角度,用可视分析技术完整地支持他们的异常检测过程.另外,面对主观性较强的可视化与交互方法研究以及领域经验要求较高的网络安全数据分析,也很少有研究者去探索指导网络安全可视化研究的模型和理论.

网络流量是主要的网络状态之一,当大部分网络故障和入侵行为发生时,往往会在网络流量中留下线索.随时间变化是网络流量的核心表征之一,从时序的角度分析网络流量是一种常见的异常检测策略,网络安全分析人员利用网络流量时序数据的多粒度、多维度和多主体性特征,可以完成对多种网络异常的发现、定位和趋势判断.

本文以网络流量时序数据的可视分析过程为研究对象,通过调研、建模、原型系统设计和案例分析,探讨了以用户分析流程模型为指导的网络安全可视化研究思路.首先,我们抽象了网络安全分析人员用网络流量时序数据检测网络异常的过程,提出了一个自顶向下的 ODSP 分析流程模型,该模型包括网络整体时序分析(overview)、网络主体时序分析(detail)、相似性网络主体分析(similarity)以及时序短期预测(prediction)这 4 个部分,它概括了分析人员由总及分、由点到面和由历史到未来的三大主要分析方向.然后,我们以 ODSP 模型为指导,综合考虑简单易用、自动化与可视化结合、界面利用率等因素,设计并实现了一个网络流量时序数据可视分析系统,该原型系统包括 4 个可协同交互的视图:时序化的平行坐标视图、多主体的矩阵视图、多主体的时序视图和相似性扩展树视图.在案例分析中,我们用原型系统分析了端口扫描和 DDoS 攻击等网络异常.最后,我们探讨了本文研究思路的优点和缺点以及我们的原型系统还有待改进之处.

1 相关工作

经过 10 多年的发展,网络安全可视化领域的学者们提出了许多新颖的可视化设计,设计并实现了诸多实用的交互式可视分析工具,比如,Zhang 等人[3]根据矩阵图、平行坐标、节点连接图等可视化技术对相关研究进行了分类,Shiravi 等人[4]根据主机监控、内外网活动监控、路由行为分析等功能分类对已有研究进行了梳理,赵颖等人[5]则从异常检测、特征分析、关联分析和态势感知等不同层次的安全需求的角度对相关研究进行了综述.

网络流量是一种随着时间产生的网络安全数据,当拒绝服务攻击和蠕虫传播等网络入侵行为发生时,往往会在网络流量上出现明显的变化,因此,很多研究者以网络流量时序可视化作为异常检测的切入点.网络流量时序数据最常见的展示方法是单个或一组时间线.为了在同一时间线中表示多个网络流量统计对象随时间变化的情况,堆叠柱状图和堆叠面积图是常用的可视化图形,比如,Abdullah 等人[6]和 Yegneswaran 等人[7]分别用堆叠柱状图和堆叠面积图可视化多个监控端口的流量变化情况;Zhao 等人[8]将两组堆叠面积图沿中轴分别向上和向下堆叠,从而提高对比分析两组时序的效率.另一种常见的网络流量时序数据可视化方法是带有时间维度的二维点阵图(矩阵图或像素图),一个维度为时间,另外一个维度用来表示不同的主机或端口等统计对象,时序值用颜色或者形状表示,这种图形适合从整体上同时展示更多的时序信息[9−11].除了上述几种经典方法,研究者们也一直在探索新颖的网络流量时序数据可视化方法,比如,CCSvis[12]用三维圆柱体的形式来分析域名服务器的访问流量时序数据;ClockEye[13]用圆上的 24 个扇片表示一台主机或者一个子网一天的流量变化情况,大量 ClockEye 用类似 Treemap 的方法进行层次化布局,帮助网络安全分析人员自由观察 24 个小时内全网、局部子网和主机这 3 个层次的网络流量情况.

除了对网络流量时序数据进行直接的可视化编码以外,许多智能化的方法也与之结合起来.比如,TVi[14]设计了一个结合信息熵与主成份分析的异常检测器,并将其与网络流量时序可视化工具结合起来检测网络异常;Stoffel 等人[15]将小波分析与时序可视化结合起来,寻找具有相似行为的主机.随着网络规模的不断扩大,研究者们开始逐步深入地探讨如何快速处理大规模流量时序数据,以及如何可视化大规模网络监控对象的流量变化.比如,NStreamAware[16]设计了一个基于 Spark 的网络流量实时处理平台,用于支持他们设计的 NVisAware 网络安全态势可视化工具;SpringRain[17]提出了一种以瀑布为隐喻的大规模网络态势可视化设计思路,大规模网络监控对象的流量和其他状态像瀑布中数以万计的水滴一样顺着时间从上往下流淌.近几年,可视分析技术逐渐兴起,各种描述网络流量和其他网络状态的时序图形,在一定程度上已成为各种可视分析工具不可或缺的标准模块,经典的设计模式如 OCEANS[18],它用网络流量和报警次数等多组时间线刻画网络整体态势情况,分析人员可以选中关心的时段,在其他几个协同的视图中分析该时段的具体信息.总的来说,网络流量时序数据的分析与可视化在网络安全研究中扮演了重要的角色,也取得了一些阶段性成果,然而,很少有研究者去抽象和总结分析人员分析网络流量时序数据的常规流程,并用可视分析的思想支持基于网络流量时序的网络异常检测过程.

2 基于 ODSP 的网络流量时序数据分析流程

从统计意义上讲,时间序列就是将某一个指标在不同时间上的不同数值,按照时间的先后顺序排列而成的数列.网络流量监控数据可以分为比特级别、包级别和流级别这 3 种监控粒度,通常包含的数据属性有源 IP 地址、源端口号、目标 IP 地址、目标端口号、协议、上行流量、下行流量等.依据时间序列的定义,网络流量数据中的某些属性按照一定时间粒度进行聚合,就可以得到典型的时序数据.比如,整个网络内的连接次数以小时为时间单位进行聚合,一天内可得到长度为 24 的网络连接次数时间序列;又如,某个服务器的流量以分钟为时间片进行统计,1 小时可得到长度为 60 的流量大小时间序列.因此,网络流量时序数据有多层次粒度、多时间粒度、多维统计属性和多分析主体等几大特点.

为了在设计可视分析工具时遵循分析人员对网络流量时序数据的分析流程,我们调研了网络安全分析人员对网络流量时序数据的异常检测过程,收集了他们的分析技巧.通过整理和总结,我们抽象出一个典型的自顶向下的网络流量时序数据分析流程,称为 ODSP 模型,如图 1 所示.

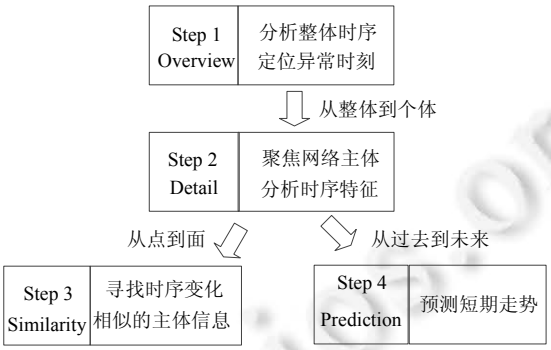

Fig.1 Flow diagram of the analysis of network traffic time-series based on ODSP model

图 1 基于 ODSP 的网络流量时序数据分析流程图

- 第 1 步,通过分析整体时序(overview)定位异常时刻:以整个网络或者某个范围内的子网为分析对象,选取多个特征构造时序数据,比如网络总体流量大小和连接次数、活跃的主机数和网络协议数等,根据领域知识定位到感兴趣时刻,比如寻找流量较大或较小的时刻、寻找连接次数大但活跃主机数少的时刻.
- 第 2 步,通过聚焦网络主体(detail)分析异常主机或端口的时序特征:网络的整体流量是由网络主体的

活动形成的,主机是网络活动的首要主体,端口则标识了不同的网络应用.因此,在第 1 步中定位的流量异常时刻通常是由一些主机和端口的异常行为造成的.在第 2 步中,首先要快速聚焦到在这些时刻有可疑行为的主机和端口,为了进一步确认它们的可疑行为,还需要观察一段时间内它们的流量时序特征,比如某 Web 服务器 80 端口是否存在周期性的流量高峰、某主机的 6667 号端口是否在此之前从来没有过活动.

- 第 3 步,寻找时序变化相似的主体信息(similarity):复杂的网络安全事件通常都具有多步性和协作性的特点,在第 2 步中找到的有异常行为的主机和端口是很好的参照物,具有相似行为模式的其他主机和端口很可能是协同攻击者或者其他受害者.比如,发起 DDoS 攻击的协同攻击者的流量变化情况,在攻击期间会表现出较高的相似性.
- 第 4 步,预测流量的短期走势(prediction):前 3 步是从历史流量时序数据中定位异常时刻和异常主体,而分析人员在选择异常处理方案时,则需要预判未来的走势.因此,第 4 步的趋势预测是根据历史数据和当前迹象推测网络的短期走势,为分析人员的决策提供辅助信息.

ODSP 是以流量时序数据为中心的分析流程,融合了分析人员通过流量时序数据检测网络异常的几种常见的分析思路:(1) 由整体到个体,通过整体流量变化的线索,逐步找到异常的网络主机和端口;(2) 由点到面,通过可疑的主机和端口的流量变化特征,寻找相似趋势的其他主机和端口;(3) 由过去到未来,通过历史流量变化特征预测短期走势,提供决策参考.但在实现可视化支持的 ODSP 分析流程时也面临着很多问题,下面我们具体介绍本文的原型系统设计.

3 原型系统设计

基于 ODSP 分析流程,本文设计的原型系统包括 4 个协同交互的可视化视图和一个控制面板,如图 2 所示.

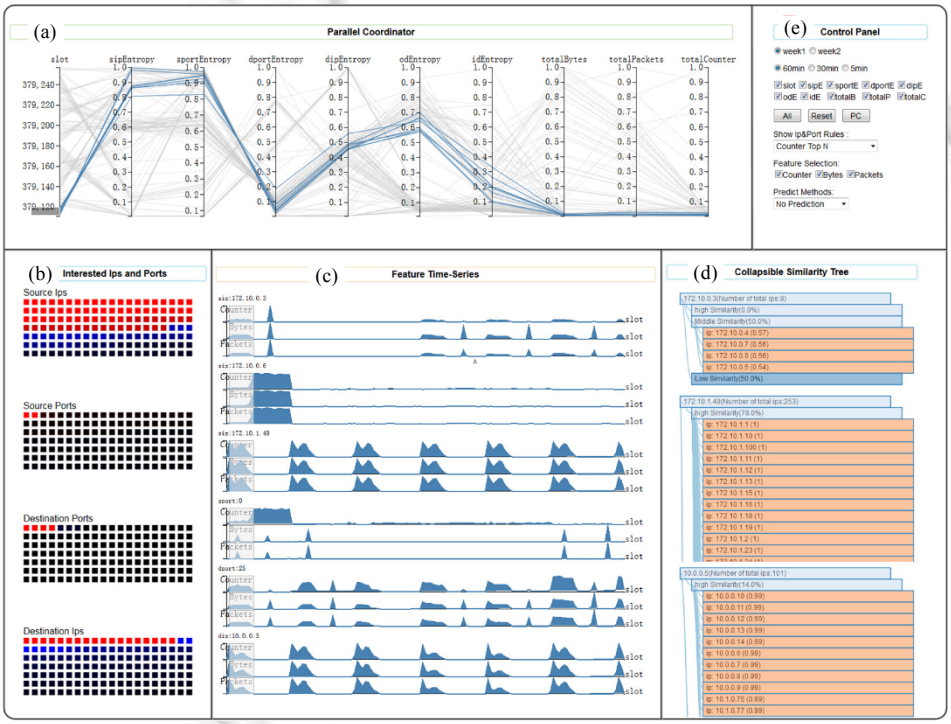

Fig.2 Prototype system interface overview

图 2 原型系统界面概览图

4 个可视视图分别是:时序化的平行坐标视图、多主体的矩阵视图、多主体的时序视图和相似性扩展树视图.平行坐标视图对应网络整体流量时序的多维特征分析,矩阵视图对应多种网络活动主体在某时段的细节分析,时序视图对应多种网络活动主体的时序特征细节分析和短期预测,相似性扩展树视图用于寻找具有相似行为特征的网络活动主体.

3.1 时序化的平行坐标视图

在 ODSP 分析流程中,网络流量整体时序分析的主要难点在于多维性,这又包括两个方面:(1) 特征选择,选择哪些特征来分析网络整体流量;(2) 多维时序可视化,如何可视化多特征的流量时序数据.

网络流量日志是一个典型的多维数据集,以流级的 Netflow 监控数据为例,主要维度包括源 IP 地址、目的 IP 地址、源端口号、目的端口号、协议类型、连接时长、包数等.除了使用传统的连接次数、字节数、包数、协议数等统计特征以外,本文还使用了以信息熵为基础的 6 个流量特征:源 IP 熵、目的 IP 熵、源端口熵、目的端口熵、出度熵和入度熵.

在信息论中,熵用来度量随机变量取值的不确定性程度:取值越是有序,信息熵就越低;反之,取值越是无序,信息熵就越高.网络流量的每一个维度都可以看作一个离散随机变量,在网络安全领域,有大量的相关研究表明[19,20]:用熵作为流量特征可以有效地度量网络活动的随机特征,从而帮助分析人员快速地进行异常检测.下面我们以目的端口熵为例介绍熵值计算方法.设流量日志中出现的目的端口号为离散随机变量 $\boldsymbol{Y}$,那么 $\boldsymbol{Y}$ 的取值空间为 $\boldsymbol{Y}=\{y_j,j=1,\ldots,m\}$,$m$ 是端口号,其实际取值范围为 0~65 535,如果在某时间段内,目的端口号 y_j 出现的概率表示为 p_j,则该时间段内网络的目的端口熵可以表示为

$$H(\boldsymbol{Y})=-\sum_{j=1}^{m}p_j\log p_j \tag{1}$$

如果目的端口熵值 $H(Y)$越大,表示活动的目的端口越随机,也就是说,有更多的端口以较为相近的概率出现;相反,熵值$H(Y)$越小,表示活动的目的端口越有序,也就是说,只有少量的活动端口.因此,目的端口熵$H(Y)$提供了对一段时间内的网络活动中目的端口的随机特征的描述,这种描述可以用来检测网络异常.比如,当恶意端口扫描发生时,网络中会出现大量的探测数据包,这些数据包可能会集中访问目的主机的绝大多数端口,并且对每个目的端口探测的概率会相对平均,此时,目的端口熵会很高;当 DDoS 攻击发生时,为了让目的主机的某些服务陷入瘫痪,攻击者常常会发送海量的攻击包到目的主机的固定端口,被攻击端口出现的概率会大幅增加,这时,目的端口熵会突然变小.

分析人员需要从多个角度把握网络流量,对可视化的需求必然是希望同时观察多特征的整体流量时序,但传统的时间线组过于占用空间,带时间维度的像素图难以辨识具体时序值且交互性差.另外,对分析人员而言,最重要的操作需求是结合自己的专业知识,通过快速的多条件过滤来定位异常的时间段.平行坐标轴是一种常见的多维数据可视化技术,它能够清晰地展示各维度的取值及其分布,能够支持多维协同快速过滤,且较为节省屏幕空间.因此,本文用平行坐标轴来可视化多特征的流量时序.为了解决平行坐标轴的时间辨识问题,我们将时序上每个时间片进行编号,将时间片编号作为一个标识维度加入平行坐标轴,这样就形成了如图 3 所示的时序化的平行坐标轴.图 3 左边是 6 个时间线表示的网络整体流量时序数据;图 3 右边是对应的时序化平行坐标轴,其中时间片编号是新增维度,图中展示了 3 个连续时间片在两种图形上的可视化效果.另外,各个流量特征具有不同的量纲和数量级,我们对数据进行标准化处理,除了时间维度以外,对其他维度上的取值均进行除以最大值的归一化处理,使得取值都落在 0 与 1 之间.

在图 2(a)所示的平行坐标视图中,交互操作主要包括:增加和删除维度、移动轴的位置、在多轴上选择感兴趣区域来过滤和高亮数据;另外,在控制面板中有 60 分钟、30 分钟和 5 分钟这 3 种时间粒度可供选择.当分析人员完成数据过滤后,满足过滤条件的记录对应的时间片将作为参数传入矩阵图.在矩阵图中,我们将通过自动化与可视化结合的思路,帮助分析人员找出在这些选中时刻中值得关注的网络主机和端口,从而由整体分析转入网络活动主体的细节分析.

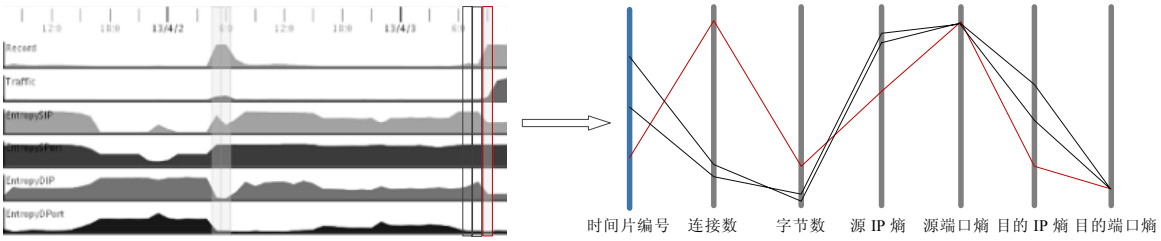

Fig.3 Illustration of parallel coordinate with temporal feature

图 3 时序化平行坐标轴的示意图

3.2 多主体的矩阵视图

在时序化的平行坐标视图中,分析人员从网络整体流量的角度定位到异常时刻,在获得部分感兴趣的时刻之后,需要对这些时刻的详细信息进行展示,为聚焦到具体的主机及端口做准备.主机作为网络活动的主体,是最重要的网络监控对象,不同端口则关联了具体的网络应用.因此,从整体流量分析转入个体流量分析时,我们着重关注源 IP 地址、源端口、目的端口、目的 IP 地址的流量信息.

像素图和 Treemap 等许多可视化方法都曾被用于展示某时段网络流量在主机和端口上的分布情况,但成千上万的主机和端口会占用较大的屏幕空间.另外,传统方法只是将流量在大量主机和端口上的分布可视化出来,完全由分析人员的视觉辨识能力去寻找异常主机和端口.因此,为了在有限的屏幕空间中展示 4 种网络主体,并帮助分析人员更快速地定位最可能的异常网络主体,本文采用了将自动化与可视化的方法结合起来的思路.首先,我们提供了一些自动化的过滤器;然后,将符合过滤条件的源 IP 地址、目的 IP 地址、源端口、目的端口用矩阵的形式可视化出来.目前,原型系统中提供的过滤器包括:流量 Top-*N*、连接数 Top-*N*、活跃端口数 Top-*N*、活跃 IP 数 Top-*N*.在图 2(b)所示的矩阵视图中,4 种网络活动主体分 4 块进行布局,每一个小方格子代表一个 IP 地址或者端口号,其颜色根据所选过滤器对应的值进行编码,由大到小渐进地从暖色系向冷色系映射.比如,颜色越红代表主体越活跃,蓝色代表主体活跃度适中,黑色代表主体极度不活跃.

3.3 多主体的时序视图

矩阵视图只展示了网络活动主体在特定时段内的流量信息,对于感兴趣的 IP 地址和端口,分析人员需要观察一段时间以来这些 IP 地址和端口的流量变化情况,以进一步确定它们的行为特征.因此,当某些网络活动主体在矩阵视图中被选中时,它们对应的时序数据将显示在时序视图中.

时序视图采用时间线组的形式,如图 2(c)所示,支持多种网络主体和多特征时序的同时展示,这有利于分析人员从复杂的对比分析中获得信息.比如,分析人员可以查看多个 IP 地址和端口的时序,并可以同时观察它们的连接数和字节数等流量特征的时序.在目前的原型系统中,在时序视图中可以添加的网络活动主体包括 IP 和端口,可以使用的流量特征包括连接数、字节数和包数.

时序视图整合了短期预测的功能,用于给出时序视图中各时序的短期预测值,供分析人员在决策时参考.时间序列预测是时间序列分析中的一个重要研究领域,有很多预测模型和方法.我们观察并分析了网络主机和端口产生的流量时序特点,将其归为 3 种类型的时序:(1) 第 1 种是在常数值附近波动的平稳时间序列数据;(2) 第 2 种是表现出一定趋势的非平稳时间序列数据;(3) 第 3 种是带有周期性的非平稳时间序列数据.

针对 3 种类型的流量时序,我们引入了 3 种经典的预测模型对网络主机和端口产生的流量进行预测,它们分别是:对平稳序列进行预测的自回归移动平均模型(ARMA-autoregressive moving average model)、对非平稳序列进行预测的差分自回归移动平均模型(ARIMA-autoregressive integrated moving average model)、对周期性非平稳序列进行预测的季节性差分自回归移动平均模型(seasonal ARIMA).

在时序视图中,分析人员可以对任意一个已展示的时序进行预测,预测模型和预测步长在控制面板中选择.图 4 是一台 IP 为 10.1.0.75 的主机 6 天的流量时序数据,这台主机表现出了一定的周期性,其中,周期为一天.但

是每天的周期性在趋势上又有细微差别,因此,我们选择 Seasonal ARIMA 对其进行预测,预测结果紧跟历史数据,用浅灰色背景凸显在时序视图中.

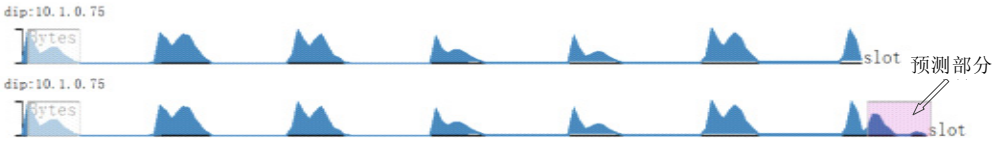

Fig.4 Prediction result of network traffic time-series

图 4 网络流量时序的预测结果

3.4 相似性扩展树视图

当出现网络攻击时,通常会有多个攻击者协同行动,而受害者也并一定是孤立的.因此,对主机和端口的行为模式进行相似性度量,能够帮助分析人员找出发起网络攻击的多个攻击者以及被攻击的多个受害对象.如图2(d)所示,相似性扩展树视图用于帮助分析人员从单个网络活动主体分析转向相似的多个网络活动主体分析.

有很多方法可以用来分析时序的相似性,本文重点从趋势和效率两个方面来设计相似性分析方法.协同攻击者和多个受害者的流量时序最重要的相似特征是同增同减的趋势,因为主要攻击者和主要受害者在攻击发生时流量会有明显的变化,很容易在前面几个视图中找到,但协同攻击者和其他受害者可能流量变化的幅度会明显减小,不易发现,但它们的趋势是相同的,所以趋势相同的时序变化是寻找这些协同攻击者和其他受害者的重要线索.另外,还需要控制计算相似性的时间消耗,因为有大量的网络主机和端口需要计算相似性.本文采用基于时序趋势的相似性度量方法,通过统计出两个时序同增同减的次数,并计算同趋势次数占总时间片的比例,得到相似度大小.假设某主机的时序为 $T(i)$,i=1,…,n,时序变化趋势计算为

$$\Delta T(i)=\begin{cases}1, & T(i)-T(i-1)>0\\ 0, & T(i)-T(i-1)=0\\ -1, & T(i)-T(i-1)<0\end{cases},\ i=1,...,n \tag{2}$$

$\Delta T(i)$是时序 $T(i)$两个相邻时刻的增量符向量,表示时序的变化趋势.通过统计两个时序 $T_1(i),T_2(i)$的增量符向量 $\Delta T_1(i)$和$\Delta T_2(i)$中相同趋势的个数在整个时序中的比例来计算趋势的相似性,时序 $T_1(i),T_2(i)$的趋势相似计算公式为

$$Sim_{1\text{ and }2}=\frac{\sum_{\cdot\geqslant 0}\Delta T_1(i)\cdot\Delta T_2(i)}{n} \tag{3}$$

其中,n 为时序的长度.另外,一次协同的网络攻击通常会在短时间内完成,为了更好地寻找趋势相同的网络活动主体,选择相对长一点的时间片单位,比如 30 分钟和 60 分钟,可以更好地实现趋势相似性的匹配.

本文设计了相似性扩展树来层次化的可视化相似性计算结果.相似性扩展树一共分为 3 级:(1) 第 1 级是分析人员在时序视图中选中的主机;(2) 第 2 级是对相似性进行分等;(3) 第 3 级显示了每个相似性等级下对应的 IP 地址.

我们对相似性进行分等是为了节省屏幕空间,同时便于分析人员快速查看高度相似的网络活动主体.本文的相似度被分成 3 等,分别定义为高相似性、中等相似性以及低相似性.分析人员可以设置其阈值,比如,高相似性为大于 80%,中等相似性为 50%~80%,低相似性为 20%~50%.在第 2 级的相似性等级边上,还会显示有多少比例的主机或端口被划入该等级.图 5 显示了 IP 地址为 172.10.0.5 的相似性扩展树,其中,IP 地址 172.10.0.4 与其相似度达到了 95%,分析人员可以点击这个高相似 IP 地址,在时序视图中比较它们的细节.

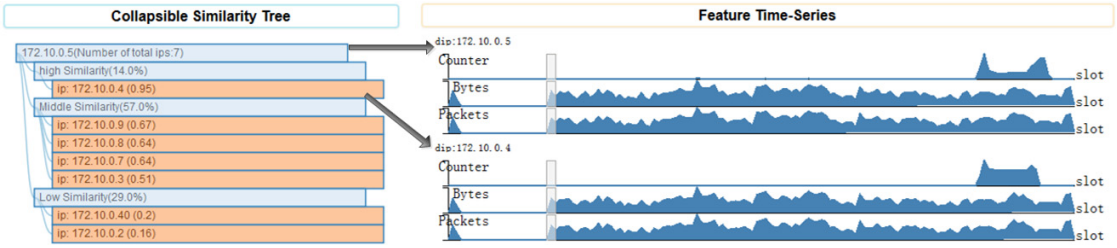

Fig.5 View of collapsible similarity tree for network traffic

图 5 网络流量的相似性扩展树视图

4 案例分析

本文选取的实验数据为可视分析挑战赛 VAST 2013 Challenge-Mini Challenge 3[21]提供的 Netflow 日志,该流量监控日志时间跨度为两周,约 8 000 万行记录,包括约 1 100 台主机及服务器.在案例分析中,我们根据 ODSP 的流程,用原型系统检测并分析实验数据中潜在的端口扫描和 DDoS 攻击.

4.1 端口扫描的分析

以攻击者发送探测帧扫描目标主机的所有端口为例,这种端口扫描通常会表现出非常高的目的端口熵,因为大量探测帧会平均地访问目标主机的大部分端口.于是,我们在整体流量时序分析时选取了源端口熵、目的端口熵、总字节数、总包个数、总连接数这 5 个特征进入平行坐标.如图 6(a)所示,我们很轻易地找到了目的端口熵接近 1 的一个时刻:2013 年 4 月 6 日 11:00,通过高亮该时刻,该时刻的其他几个流量特征值也清晰可见,源端口熵不大,说明参与扫描的源端口并不多,流量相对较大,而连接数不高.

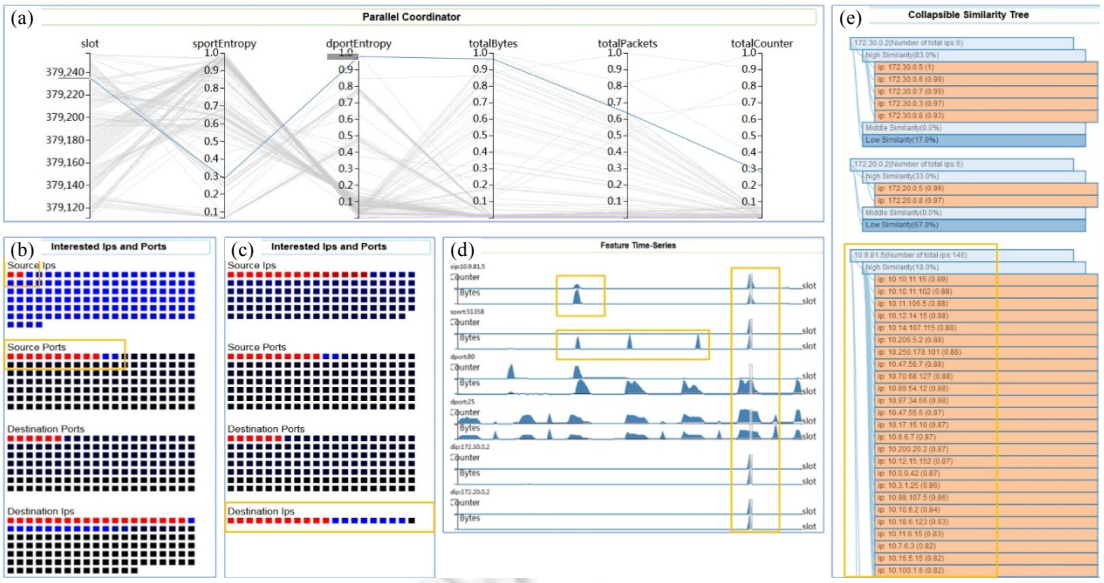

Fig.6 Visual analysis of a port scan event

图 6 端口扫描事件的可视分析

为了快速聚焦到可能发起扫描的主机,我们选择流量 Top-140 作为过滤器,在矩阵中显示这个时刻流量较大的源、目的主机以及端口.如图 6(b)所示,两台高亮的源主机 10.9.81.5 和 10.10.11.15 以及约 10 个非常规端口(51358,51357,45032 和 62559 等)引起了我们的注意.为了进一步确定被扫描的目的 IP 地址,我们修改过滤器为

活跃端口数 Top-140,如图 6(c)所示,目的 IP 矩阵发生了明显的变化,少量活跃端口数非常大的目的 IP 地址凸显出来,这几个 IP 地址是网络中的几台 Web 和邮件服务器,它们的 6 万多个端口被访问了.

为了观察攻击者和受害者在其他时间的流量情况,我们选出攻击者 10.9.81.5、发起攻击的源端口 51358、受害服务器 172.30.0.2 和 172.20.0.2 进入时序视图进行分析,如图 6(d)所示.两个受害服务器在最近一周中只出现了一次流量高峰,但攻击者 10.9.81.5 却还有一个流量高峰出现,而且发起攻击的源端口 51 358 的流量有周期特征,这些线索将引导发现可能存在的其他异常.

在相似性扩展树中,我们寻找与时序视图中的 3 个 IP 的具有相似时序特征的主机,图 6(e)显示具有高相似度的受害者并不多,而且在矩阵图中都已经被发现,但具有高相似度的攻击者却非常多,这说明同时发起目的端口扫描的源主机远比矩阵图中发现的两个高亮源主机要多.

综上所述,2013 年 4 月 6 日 11 点左右,10.9.81.5 等许多外部主机借助几个不常用的端口,对监控网络的几台 Web 和邮件服务器进行了目的端口扫描,约 6 万多个目的端口被扫描了,我们推测,目的是寻找潜在的漏洞端口,企图利用漏洞对这些服务器进行后续的攻击.

4.2 DDoS攻击的分析

如果说端口扫描只是攻击准备阶段,那么 DDoS 攻击则是一种危害极大的常见网络攻击.DDoS 攻击是指将多个攻击者联合起来,对一个或多个目标对象在短时间内发起海量的恶意访问请求,并导致目标对象的某些网络服务无法正常工作甚至直接宕机.如图 7(a)高亮了一个明显的 DDoS 攻击时刻,2013 年 4 月 11 日中午 12:00.从平行坐标轴来观察这个时刻的网络整体流量特征,连接数和包数都非常大.目的端口熵接近 0,说明可能被攻击的目的端口很少;目的 IP 熵不大,说明可能被攻击的目的 IP 数有多个;源 IP 熵不大,说明攻击者可能较多.

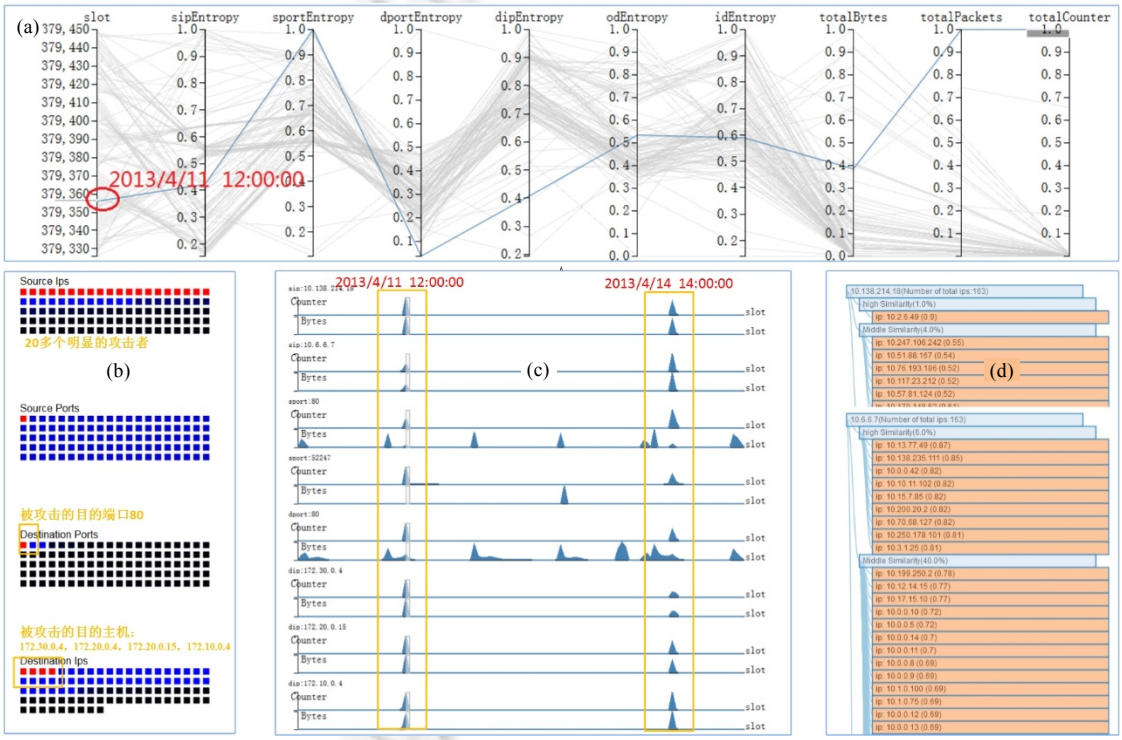

Fig.7 Visual analysis of a DDoS attack

图 7 DDoS 攻击的可视分析

为了快速定位攻击者和被攻击者,我们在矩阵图中选择了连接数 Top-100 观察这个时刻活跃的网络主机和

端口.如图 7(b)所示,在源端口矩阵中,有 20 多个明显的攻击者,被攻击的端口集中在 80,被攻击的目的主机是 4 台 Web 服务器.当我们选择了端口 80、几台攻击主机和被攻击的 Web 服务器并观察它们近一周的流量走势时,在图 7(c)中发现,在 2013 年 4 月 14 日 14:00 也出现了同样的流量峰值.在后续分析中我们发现,这是另外一次 DDoS 攻击,但攻击的量并没有 4 月 11 日中午那次大.而我们选择了几个攻击者,并在相似性扩展树中观察与它们流量时序相似的主机,如图 7(d)所示,其他大量潜在的协同攻击者出现在了扩展树中.

5 总结与展望

本文尝试了一种以分析流程为中心的网络安全可视化研究思路:首先,本文提出了一个 ODSP 的网络流量时序分析流程模型,该模型自顶向下地融合了分析人员由整体到个体、由点到面和由历史到未来的网络流量时序数据分析和异常检测过程;然后,以该模型为指导,本文设计并实现了一个多视图协作的可视分析原型系统,用于帮助分析人员实现可视化的网络流量时序数据分析过程.在案例分析中,以最常见的端口扫描和 DDoS 攻击为例,我们的原型系统从网络整体流量分析、多网络主体的流量分析、相似网络主体搜索和流量趋势短期预测这 4 个方面,较好地支撑了分析人员从网络流量时序数据中检测网络异常的整个过程.

这次有益的尝试,为我们的研究积累了经验,也为后续研究提出了更多的思考.本文的原型系统目前还处于内部测试阶段,并没有应用到实际网络环境中,还需要大量的工作来提高系统的整体运行效率、可用性和鲁棒性.本文强调以分析人员的日常分析流程为中心,在设计中尽可能地使用简单易用的可视化技术,但很多可视化设计的细节还有待改进,信息的可视化传递还可以更为丰富和细腻.本文的原型系统中设置了很多可以结合智能化方法的接口,比如网络主机和端口的过滤器、基于信息熵的流量特征提取等,在后续研究中,还将对这种互相取长补短的设计思路进行更多的尝试.本文选用了一种自顶向下的分析流程作为参照模型,但不同分析人员的背景知识和习惯各不相同,即便都是针对网络流量时序数据,分析人员也可能自底向上地分析,比如,先掌握核心路由器和重要服务器的流量变化,然后推演对其他主机或网络整体的影响.因此,网络安全数据分析流程的多样化,将为可视分析技术的应用带来机遇和挑战.

References:

[1] Lü LF, Zhang JW, Sun JZ, He PL, Sun LG. Survey of network security visualization techniques. Journal of Computer Applications, 2008,28(8):1924−1927 (in Chinese with English abstract). [doi: 10.3724/SP.J.1087.2008.01924]

[2] Harrison L, Lu AD. The future of security visualization: Lessons from network visualization. IEEE Network, 2012,26(6):6−11. [doi: 10.1109/MNET.2012.6375887]

[3] Zhang YP, Xiao Y, Chen M, Zhang JY, Deng HM. A survey of security visualization for computer network logs. Security & Communication Networks, 2012,5(4):404−421. [doi: 10.1002/sec.324]

[4] Shiravi H, Shiravi A, Ghorbani AA. A survey of visualization systems for network security. IEEE Trans. on Visualization and Computer Graphics, 2012,18(8):1313−1329. [doi: 10.1109/TVCG.2011.144]

[5] Zhao Y, Fan XP, Zhou FF, Wang F, Zhang JW. A survey on network security data visualization. Journal of Computer-Aided Design & Computer Graphics, 2014,26(5):687−697 (in Chinese with English abstract).

[6] Abdullah K, Lee C, Conti G, Copeland JA. Visualizing network data for intrusion detection. In: Proc. of the 6th Annual IEEE SMC Information Assurance Workshop. New York: IEEE SMC, 2005. 100−108. [doi: 10.1109/IAW.2005.1495940]

[7] Yegneswaran V, Barford P, Ullrich J. Internet intrusions: Global characteristics and prevalence. In: Proc. of the 2003 ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. New York: ACM Press, 2003. 138−147. [doi: 10.1145/781027.781045]

[8] Zhao Y, Liang X, Fan XP, Wang YW, Yang MJ, Zhou FF. MVSec: Multi-Perspective and deductive visual analytics on heterogeneous network security data. Journal of Visualization, 2014,17(3):181−196. [doi: 10.1007/s12650-014-0213-6]

[9] Berthier R, Cukier M, Hiltunen M, Kormann D, Vesonder G, Sheleheda D. Nfsight: Netflow-Based network awareness tool. In: Proc. of the 24th Int'l Conf. on Large Installation System Administration. USENIX Association, 2010. 1−8.

[10] Taylor T, Paterson D, Glanfield J, Gates C, Brooks S, McHugh J. Flovis: Flow visualization system. In: Proc. of the Conf. for Homeland Security. IEEE Computer Society, 2009. 186−198. [doi: 10.1109/CATCH.2009.18]

[11] Conti G, Abdullah K, Grizzard J, Stasko J, Copeland JA, Ahamad M, Owen HL, Lee C. Countering security information overload through alert and packet visualization. IEEE Computer Graphics and Applications, 2006,26(2):60−70. [doi: 10.1109/MCG.2006.30]

[12] Seo I, Lee H, Han SC. Cylindrical coordinates security visualization for multiple domain command and control botnet detection. Computers & Security, 2014,46:141−153. [doi: 10.1016/j.cose.2014.07.007]

[13] Fischer F, Fuchs J, Mansmann F. ClockMap: Enhancing circular treemaps with temporal glyphs for time-series data. In: Proc. of the Eurographics Conf. on Visualization (EuroVis). Eurographics, 2012. 97−101.

[14] Boschetti A, Salgarelli L, Muelder C, Ma KL. Tvi: A visual querying system for network monitoring and anomaly detection. In: Proc. of the 8th Int'l Symp. on Visualization for Cyber Security. Pittsburg: ACM Press, 2011. 1−10. [doi: 10.1145/2016904.2016905]

[15] Stoffel F, Fischer F, Keim DA. Finding anomalies in time-series using visual correlation for interactive root cause analysis. In: Proc. of the 10th Workshop on Visualization for Cyber Security. New York: ACM Press, 2013. 65−72. [doi: 10.1145/2517957.2517966]

[16] Fischer F, Keim DA. NStreamAware: Real-Time visual analytics for data streams to enhance situational awareness. In: Proc. of the 11th Workshop on Visualization for Cyber Security. Paris: ACM Press, 2014. 65−72. [doi: 10.1145/2671491.2671495]

[17] Promann M, Ma YA, Wei S, Lei WR, Chang JSK, Qian ZC, Chen YV. SpringRain: An ambient information display. In: Proc. of the Visual Analytics Science and Technology 2013 (VAST). Los Alamitos: IEEE Computer Society Press, 2013. 5−6.

[18] Chen S, Guo C, Yuan XR, Merkle F, Schaefer H, Ertl T. OCEANS: Online collaborative explorative analysis on network security. In: Proc. of the 11th Workshop on Visualization for Cyber Security. Paris: ACM Press, 2014. 1−8. [doi: 10.1145/2671491.2671493]

[19] Nychis G, Sekar V, Andersen DG, Kim H, Zhuang H. An empirical evaluation of entropy-based traffic anomaly detection. In: Proc. of the 8th ACM SIGCOMM Conf. on Internet Measurement. Vouliagmeni: ACM Press, 2008. 151−156. [doi: 10.1145/1452520.1452539]

[20] Zhao Y, Zhou FF, Fan XP, Liang X, Liu YG. IDSRadar: A real-time visualization framework for IDS alerts. Science China Information Sciences, 2013,56(8):1−12.

[21] VAST challenge homepage. 2013. http://www.vacommunity.org/VAST+Challenge+2013

附中文参考文献:

[1] 吕良福,张加万,孙济洲,何丕廉,孙立刚.网络安全可视化研究综述.计算机应用,2008,28(8):1924−1927. [doi: 10.3724/SP.J.1087.2008.01924]

[5] 赵颖,樊晓平,周芳芳,汪飞,张加万.网络安全数据可视化综述.计算机辅助设计与图形学学报,2014,26(5):687−697.

赵颖(1980－),男,湖南益阳人,博士,副教授,CCF 专业会员,主要研究领域为可视化,可视分析.

吴青(1983－),女,讲师,主要研究领域为高性能计算,计算机网络.

王权(1991－),男,硕士生,主要研究领域为可视分析.

张胜(1980－),男,博士,系统分析师,CCF学生会员,主要研究领域为网络与信息安全.

黄叶子(1991－),女,硕士,主要研究领域为网络安全可视化.