doi:10.3772/j.issn.1002-0470.2014.09.004

信息熵时序和树图用于 NetFlow 可视化的研究①

张 胜②*** 施荣华* 赵 颖* 周芳芳*

(*中南大学信息科学与工程学院 长沙 410083)

(**湖南商学院现代教育技术中心 长沙 410205)

摘 要 针对 NetFlow 日志规模日益扩大、变化日益加快,致使管理和分析难度日益增大的趋势,根据网络安全可视化的思想,构建了一种用时间序列(Time series)图和树图(TreeMap)相结合的方式分析 NetFlow 日志的可视化系统(简称 2T 图系统),用以快速、有效地识别网络中的攻击和异常事件,掌握网络安全态势。该系统重点考虑了 NetFlow 日志中六个特征维的信息熵,通过构建时间序列图来从宏观上掌控网络状态,同时引入树图来深度挖掘入侵细节。系统还通过创建图像特征规则,从图像上直观分析攻击,发现感兴趣的模式。通过对 VAST Challenge 2013 年网络安全可视分析竞赛数据进行分析,证明该系统可以直观地从宏观和微观两个层面感知网络安全状态,有效地识别网络攻击和辅助分析人员决策。

关键词 网络安全可视化, 安全态势评估, NetFlow, 信息熵, 时间序列图, 树图

0 引 言

随着网络应用的不断扩大,对网络安全的要求越来越高,网络信息系统安全面临严峻挑战[1,2]。传统网络安全分析依靠的现有网络安全产品面对网络数据量的急剧增大、攻击类型的急剧增多,其检测性能显得明显不足,用其分析大量的日志信息来发现网络异常的方式已不再有效。在这种情况下,产生了适应网络安全要求的网络安全可视化技术。

网络安全可视化是一个新兴的多学科融合的研究领域。网络安全可视化技术是网络安全态势感知与可视化技术的结合,它利用人类视觉对模型和结构的获取能力,将抽象的网络和系统数据以图形图像的方式展现出来,能帮助分析人员快速准确的分析日志文件,从中识别出网络状态与异常等,并且预测网络安全态势,它是解决网络信息安全问题的一种重要手段。本文试图从 NetFlow(NetFlow 是一种网络负载监控技术,能提供详细网络流量信息[3],可用于网络监控、应用监控、主机监控、安全监控以及计费等业务[4])分析入手,采用可视化方法来模拟、表示和分析网络,将人类模式识别能力强的能力引入网络安全化系统中来,使普通用户无需高级的技能就能够目睹、探索以至立即理解大量的网络安全信息,超越了传统技术高度。

1 相关工作

NetFlow 日志大小与网络规模、负载承受成正比,当发生网络安全事件时,管理及分析人员很难在海量日志中快速发现问题,更谈不上及时分析处理。因此,国内外研究者一直都在不断探索如何通过新方法新技术分析枯燥的 NetFlow 日志文件,掌握整个网络的负载及其安全运行状况。例如:Zhang 等[5]使用统计分析技术来反映网络状态;Hsiao 等[6]采用空间时间聚合技术来分析恶意网站;Yin 等[7]采用动态熵技术来检测拒绝服务(DOS)攻击;Sperotto 等[8]通过分析时间序列(time series)来进行入侵检测;Francois 等[9]采用 PageRank 技术来检测僵尸网络。

可视化方法是解决分析海量 NetFlow 日志的分析困难的一种重要可行方案,主要运用的可视化技

① 国家自然科学基金(61103108, 61402540)和湖南省科技计划博士后专项与中南大学博士后启动资金(2012RS4049)资助项目。

② 男,1975 年生,博士生,系统分析师,CCF 会员;研究方向:网络信息安全,计算机支持的协作学习,网络软件研究与应用等;联系人,E-mail: 48209088@qq.com

(收稿日期:2014-02-19)

术有散点图、平行坐标、树图(TreeMap)等。主要针对的应用有主机监控、网络监控、端口分析、攻击模式等,如:NVisonIP[10],采用散点图技术,主要用于监控内外网主机,把 1 个 B 类网络组成为 256 * 256 矩阵,用颜色来表示主机状态,目的是提高安全分析人员的态势感知能力,将来努力的方向是通过自动创建符号规则来发现新的攻击和感兴趣的模式;NetBytes Viewer[11],采用三维散点图技术,主要用于监控端口活动,通过选择观察不同时间段内每端口的数据负载,帮助用户从不同的角度观察数据和发现异常,缺点是数据量多时容易造成图像拥挤难以分析;NFlowVis[12],采用树图技术,主要用于检测攻击模式,通过将待观察网络映射到树图,可疑外部主机排列在树图周边,用于发现大型的分布式攻击,不足之处是提供的攻击信息有待进一步完善和丰富;TVi[13],采用节点链接技术,用高层来显示整个网络状态,用低层来显示选中的主机的异常现象和攻击路径,提高网络管理员揭示隐含模式的能力,不足之处是缺乏对实际运营网络检测能力的测试;Flow-Inspector[14],采用直方图、力引导图、辐射图、蜂窝图等技术,实时显示网络流数据,用于检测短期和长期的网络问题;NetSecRadar[15],采用辐射图技术,圆环上分布不同的安全事件,圆内表示观测网络中主机,用于帮助用户快速识别异常、发现攻击模式和分析事件关联,主要的缺陷是如果观测主机过多,圆内主机分布拥挤,造成安全事件难于分析。

综上所述,现有可视化系统能够直观地帮助网络管理人员快速分析 NetFlow 日志,发现网络异常,掌握网络运行状态,但在加强用户体验、提高可扩展性、降低图像闭塞性、增加态势感知能力等方面需要进一步完善。

2 系统可视化框架

本研究考虑了现有可视化系统的不足和优势,采用 2T 图——时间序列(time series)图和树图(TreeMap)的结合来分析 NetFlow,构建一个 2T 图可视化系统的三层分析框架,见图 1。一层利用时间序列图(针对网络流的 6 个主要特征的信息熵)把握整个网络的安全态势;二层用整网树图把握某个时间窗口内的网络状态;三层通过树图向下钻取分析局部细节。这样做的优点是可以方便地从宏观和微观两个层次把握网络安全,降低图像闭塞性,增加态势感知能力。

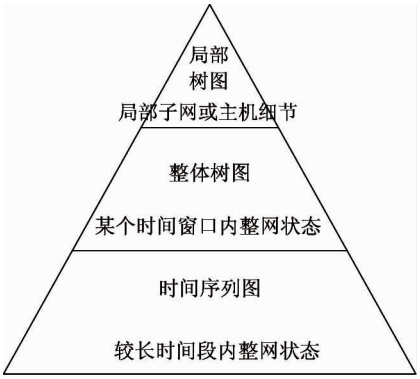


图 1 三层分析框架

2.1 时间序列设计

时间序列图用线段将各个观察数据点连接起来,对于多维数据,该图将各维分离开来,对每一维数据将其上的数值连接起来,使它能直观地反映数据在各维上的变化趋势。

2.1.1 特征熵设计

对于 NetFlow 数据,笔者经过仔细斟酌,选择最能反映数据变化趋势的 6 个维度来体现网络安全趋势:

(1) 源地址(SrcIP);

(2) 目标地址(DestIP);

(3) 源端口(SrcPort);

(4) 目标端口(DestPort);

(5) 源流每包字节数(SrcBpp);

(6) 目标流每包字节数(DestBpp)。

如果直接使用 6 个流特征的统计值,无法准确体现网络异常行为在流量中的表现模式。本文使用信息熵作为度量指标,假如 X 为离散的随机变量(6 种流量属性中一种),则信息熵定义为

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

其中 $p(x_i)$ 表示观察时间段内 x_i 出现的频率。信息熵的基本作用就是消除人们对事物的不确定性。如果数据集中于一点,也就是说所有数据具有相同的值,则信息熵为 0;相反,如果数据分布很广,则信息熵很大,例如:如果恶意软件针对一个主机的全部端口进行扫描,则目标端口信息熵很大,如果针对整个网络的同一端口扫描,则目标端口信息熵很小。这里定义网络流信息熵向量为

$$E(t) = [H_t(\text{SrcIP}), H_t(\text{DestIP}), H_t(\text{SrcPort}), H_t(\text{DestPort}), H_t(\text{SrcBpp}), H_t(\text{DestBpp})] \tag{2}$$

该向量反映了时间窗口 t 内的网络状态,信息

熵向量的时间序列 $E(1),E(2),E(3),E(4),E(5)$ 等,能够反映一段时间的网络状态。为了能够区分时间窗口 t 内网络状态是否正常,我们又引入的交叉熵,定义为

$$L_\alpha(P,Q) = \frac{1}{1-\alpha}\log_2\sum_{i=1}^{n}\frac{p_i^\alpha}{q_i^{\alpha-1}} \tag{3}$$

其中 P、Q 是离散分布,p_i、q_i 是 P、Q 的分布函数。单一的信息熵只能体现某个观测点上的静态分布状况,而交叉熵不仅考虑到流量的空间分布,同时也考虑了两个不同观测点上流量的动态变化。交叉熵的一个重要的特性是 $L_\alpha(P,Q)$ 越小,则需要获得更多的信息来区分 P 和 Q。因此通过计算 NetFlow 当前观测点与正常观测点的相对信息,就可以确定当前状态偏离正常状态的程度,并判断当前状态是否正常。为了简化计算,α 取 0.5,公式可表示为

$$L_{0.5}(P,Q) = 2\log_2\sum_{i=1}^{n}(p_iq_i)^{1/2} \tag{4}$$

这里定义正常信息熵向量 E(Normal)为

$$E(\text{Normal}) = \frac{1}{n}\sum_{t=1}^{n}E(t) \tag{5}$$

其中 n 为网络流多个正常时间窗口序列,可通过学习网络正常时间窗口内信息熵值获取。如果:

$$L_{0.5}(\text{Current},\text{Normal}) > \text{threshold} \tag{6}$$

则可以判断当前网络异常。其中 Current 为当前时间窗口内网络流分布函数,Normal 为正常时间窗口内网络流分布函数,threshold 为基础阈值(threshold 取值可由实验方法确定)。

通过信息熵的设计,最终绘制出的时间序列图如图 2 所示。用同色系颜色表示同类数据维度,其中 IP 地址用红色系,端口用绿色系,每包字节数用蓝色系,灰色竖线表示时间窗口,红色竖线表示网络异常。

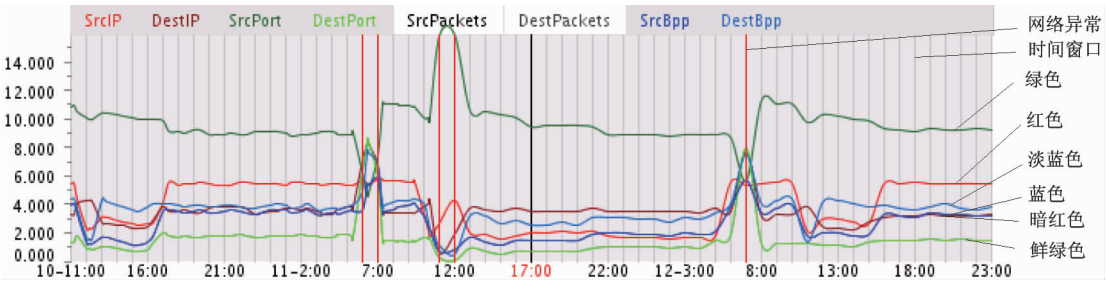


图 2　NetFlow 网络流特征信息熵图

2.1.2　异常流类型分析(图像特征)

通过分析图 2,我们可以快速感知网络安全态势,发现异常时段,但是无法区分异常流类型。下面我们提取图像特征来进行定性分析(见表 1),其中"↓"表示下降,"↑"表示增加,"—"表示变化不明显。

表 1　异常流信息熵特征

攻击类型	SrcIP 红	DestIP 暗红	SrcPort 绿	DestPort 鲜绿	SrcBpp 蓝	DestBpp 淡蓝
单目标主机多端口扫描	↓	↓	↓	↑	↓	↓
多目标主机单端口扫描	↓	↑	↓	↓	↓	↓
多目标主机多端口扫描	↓	↑	↓	↑	↓	↓
单源拒绝服务攻击	↓	↓	↓	↓	↓	↓
伪造源地址拒绝服务攻击	↑	↓	↑	↓	↓	↓
针对子网的拒绝服务攻击	↑	↑	↓	↓	↓	↓
分布式拒绝服务攻击	↑	↓	↑	↓	↓	↓
正常流	—	—	—	—	—	—

2.2　树图设计

树图是可视化层次结构数据的一个主要方法,适合于观察大数量的层次数据集,同时避免数据过度拥挤的问题。基本思想是:根据数据的层次结构将屏幕空间划分成一个个矩形(方形)子空间,子空间大小由节点大小某个特征决定,同时,对于每一个

划分的矩形可以运用其他特征进行相应的作色。树图的好处在于:更有效地利用屏幕空间降低图像闭塞性,更容易识别数据层次结构和数据节点,更易于比较数据的大小。

2.2.1　目标网络的 IP 层次规划及网络概况分析

受到桌面空间的大小限制,管理的目标主机越多,造成屏幕越拥挤。为了避免 IDSRadar[16] 出现拥挤的问题,本系统对目标 IP 采用层次结构的树图来管理,结构如图 3 所示。

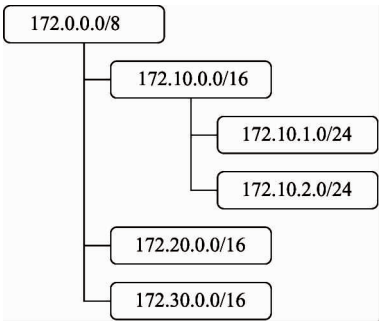
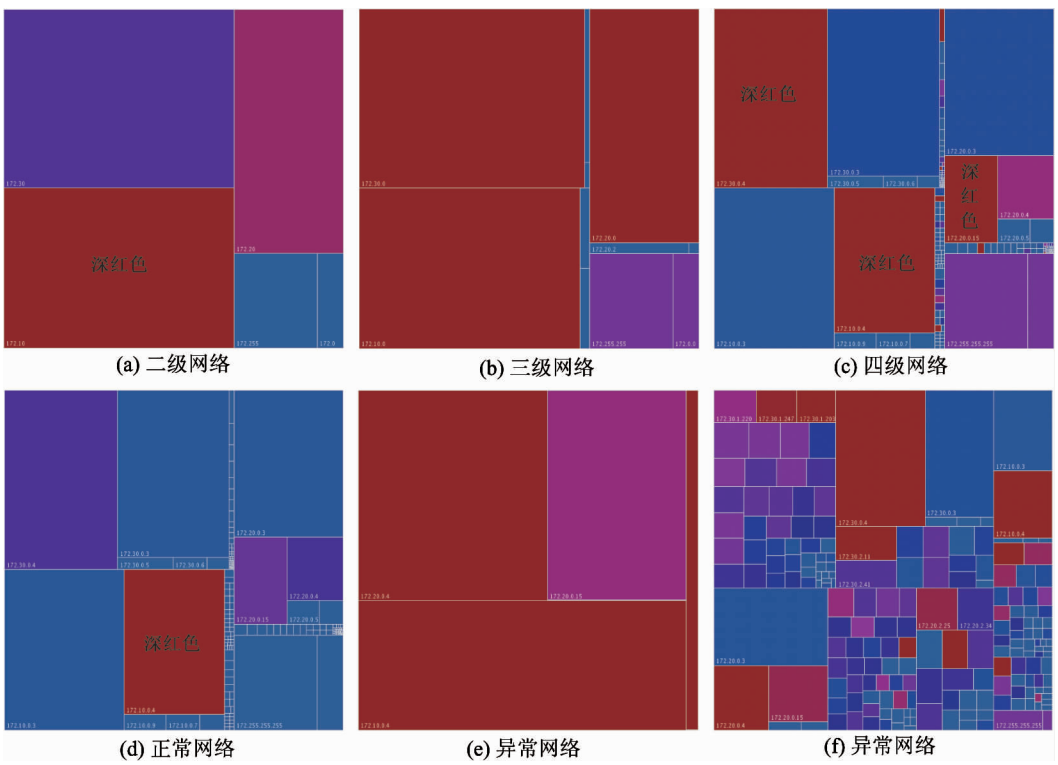


图 3　IP 层次结构图

树图中矩形尺寸和颜色可以选择表示 NetFlow 中各种特征,如:流数、流量、端口数、IP 数等。如图 4所示,矩形框表示目标 IP,尺寸表示流数(尺寸越大流数越多),颜色表示目标流量(颜色越红,流量越大)。其中图 4(a)、图 4(b)、图 4(c)分别展示了二、三、四级网络。在图 4(a)中,我们可以清楚看出流量(颜色)最大的子网是 172.10.0.0/16,流数(尺寸)最大子网为 172.30.0.0/16。在图 4(c)中,我们可以发现每个网段中都有一台主机是网络负载最重的,分别是 172.10.0.4、172.20.0.15 和 172.30.0.4。同时,树图可针对某个具体的网段作放大缩小处理,达到向下深度钻取的目的。对于某个具体的主机特征,在树图中既可以针对当前子网比较,也可以针对全网比较,如图 4(c)(局部比较)和图 4(d)(全网比较),从图 4(d)中可以直观发现全网中负载最重的主机为 172.10.0.4。



(a) 二级网络　(b) 三级网络　(c) 四级网络

(d) 正常网络　(e) 异常网络　(f) 异常网络

图 4　NetFlow 树图特征

2.2.2　异常流特征分析(图像特征)

针对异常流特征分析时,树图采用 Squarified Layout 算法绘制。网络流正常与否,取决于树图分布情况,图像过于集中或过于分散时,网络中出现异常事件的概率非常大。图 4(d)为网络流正常时的图像,图 4(e)、图 4(f)为网络异常时的图像。下面采用指数分布函数描述如下:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{7}$$

指数分布(exponential distribution)是一种连续概率

分布,用来表示独立随机事件发生的时间间隔,比如旅客进机场的时间间隔、互联网网页链接的出度入度等。经过对数据流进行统计分析,我们发现 NetFlow 符合指数分布。

如果 $\lambda >= 0.3$, 图像表示为仅少数几个矩形大块,如图 4(e)所示。形成原因:少数几台主机遭受海量网络流攻击,符合拒绝服务攻击特征;$0.1 < \lambda < 0.3$, 图像表示为矩形大块一边环绕着许多小块,如图 4(d)所示,形成原因:目标网络中存在服务器和用户机,服务器负载量较大,数据流大,树图上表现为占了屏幕中较大的空间,而其他用户的流量较小,体现为紧紧贴着服务器大区域边上的小块;$\lambda <= 0.1$, 如图 4(f)所示,图像表示为图像分布较为均匀,形成原因:目标网络中的主机遭受到的网络流攻击较为均匀,数据流数概率密度分布如图 5 所示。

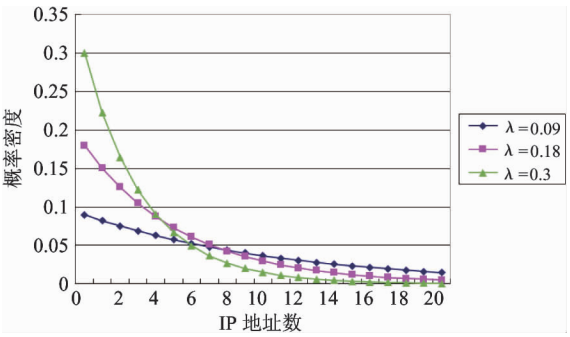


图 5　NetFlow 流数概率密度分布

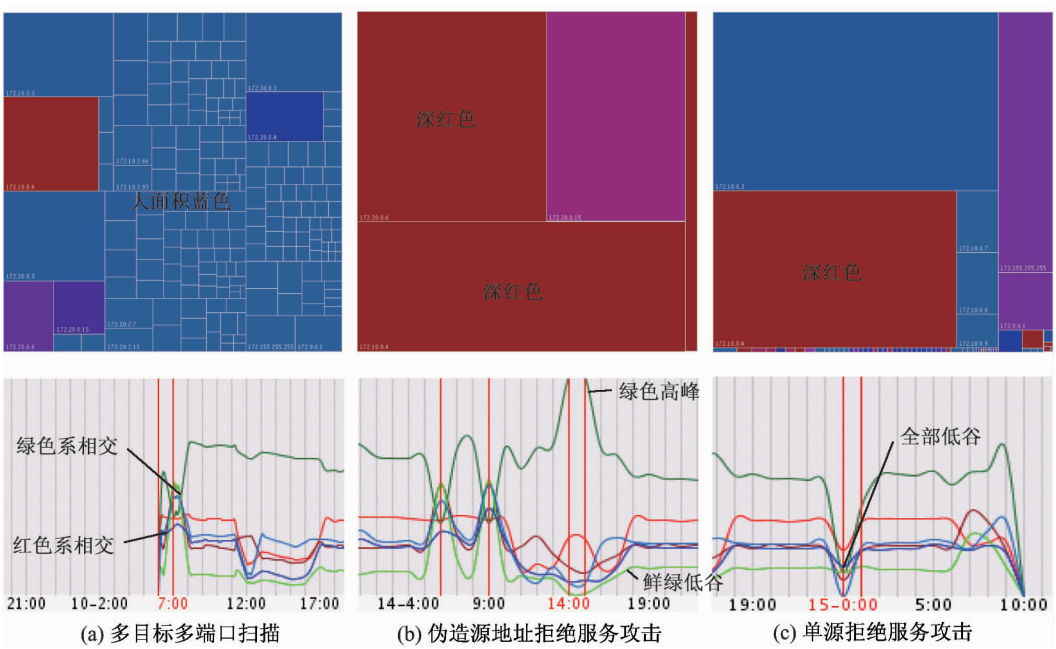
3　综合实例分析

本项目选取的实验数据来自可视化国际会议 IEEE VIS 2013 举办的可视分析挑战赛 VAST Challenge2013 年比赛数据。比赛数据提供了某跨国公司内部网络(主机和服务器约 1100 台)2300 万条的 NetFlow 日志。

通过在时间序列图上分析网络安全态势和匹配图像特征,找出网络异常时间窗口,然后在树图上进一步分析网络流特征的方法,直观快速地发现和分析问题。

图 6 为 NetFlow 异常综合分析图。如图 6(a)所示,在时间序列图中,10 日 7:00 被红色竖线标记(交叉熵阈值取 0.08),表示该处发生了网络流异常,图像特征表示为绿色系和红色系曲线各自一上一下相交。查表 1 异常类型为多目标主机多端口扫描或伪造源地址拒绝服务攻击。查看树图,矩形块分布较为均匀,绝大部分主机流量很低,可以确认是多目标主机多端口扫描。

如图 6(b)所示,14 日 14:00 被红色标记(交叉熵阈值取 0.08),绿色系线条出现一个极高峰和一个极低峰,时间序列图特征符合伪造源地址拒绝服务攻击特征。进一步查看树图,图像中只有 4 个矩形块,其中172.10.0.4和172.20.0.4承受了巨大的



(a) 多目标多端口扫描　(b) 伪造源地址拒绝服务攻击　(c) 单源拒绝服务攻击

图 6　NetFlow 异常综合分析图

网络流数据。鼠标点击 172.10.0.4 块,右边信息框中出现详细信息,该主机遭受了 14 个源地址,63344 个源端口,2 个目标端口,1441338330 字节流量的攻击;鼠标点击 172.20.0.4 块,该主机遭受了 8 个源地址,63639 个源端口,1 个目标端口,1342772662 字节流量的攻击。

如图 6(c)所示,15 日 0:00 点被红色标记(交叉熵阈值取 0.08),时间序列图所有线条都出现了低峰,符合单源拒绝服务攻击特征,查看树图,主机 172.10.0.4 颜色最红,该主机遭受的攻击流量最大。

同时,在几乎所有的树图上都发现一个有趣的模式。如图 4(c)、4(f)、6(a),右下角都有两个地址为 172.255.255.255 和 172.0.0.1 的主机,流量不大而流数不小,图像表示为两块不小的深色区域,究其原因,应为公司内部启用网络电话,这些不大的流量作为网络通话时使用的带宽。

4 系统优势

本研究构建的 2T 图可视化系统有以下优势:

(1)运用熵值表示网络安全态势,代替原有简单的信息量统计技术,采用 NetFlow6 个特征的信息熵[17,18]来绘制时间序列图,用交叉熵来区分正常流和异常流,通过简单图像匹配,分析人员能更加直观地发现网络问题,初步分析攻击类型。

(2)相对于用点阵的方式表示内网主机,本系统采用的树图层次结构来表示,有利于管理大型和超大型网络,不会出现由于显示空间造成图像拥挤,无法分辨图像的问题,起到降低图像闭塞性[19]的作用。

(3)通过一套直观图像特征分析方法和创建图像特征规则,直观分析攻击,发现感兴趣的模式。

5 结 论

可视化技术是网络安全领域一个新的研究热点。本文提出将 2T 图(时间序列图和树图相结合)用于 NetFlow 日志数据的可视分析,通过实例分析,证明本研究构建的可视化系统可帮助用户快速、直观发现网络流中异常并且分析异常类型。在今后研究中,将进一步加强可视化系统的网络态势感知能力,融合更多的可视化技术,综合分析各种安全日志(如防火墙、网络杀毒系统、主机状态系统、入侵检测系统等),借助可视化技术发挥多源安全数据综合分析的优势,帮助网络分析人员快速建立对所监管的网络整体情况的有效认知,进行态势评估,应对将来的挑战。

参考文献

[1] Cncert/Cc. 2012 年我国互联网网络安全态势综述. http://www.cert.org.cn/: 国家互联网应急中心, 2013

[2] Cncert/Cc, 2012 年中国互联网网络安全报告. 北京: 人民邮电出版社, 2013

[3] 赖积保, 王慧强, 金爽. 基于 Netflow 的网络安全态势感知系统研究. 计算机应用研究, 2007, 24(8): 167-172

[4] Li B, Springer J, Bebis G, et al. A survey of network flow applications. *Journal of Network and Computer Applications*, 2013, 36(2): 567 - 581

[5] Zhang H. Study on the TOPN abnormal detection based on the netflow data set. *Computer and Information Science*, 2009, 2(3): 103-108

[6] Hsiao H W, Chen D N, Wu T J. Detecting hiding malicious website using network traffic mining approach. In: Proceedings of the 2nd International Conference on Education Technology and Computer (ICETC), Shanghai, China, 2010. V5: 276-280

[7] Yin K X, Zhu J Q. A novel DoS detection mechanism. In: Proceedings of the 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), Jilin, China, 2011. 296-298

[8] Sperotto A, Pras A. Flow-based intrusion detection. In: Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management, Dublin, Ireland, 2011. 958-963

[9] Francois J, Wang S, Bronzi W, et al. BotCloud: detecting botnets using MapReduce. In: Proceedings of the 2011 IEEE International Workshop on Information Forensics and Security (WIFS), Foz do iguacu, Brazail, 2011. 1-6

[10] Lakkaraju K, Bearavolu R, Slagell A, et al. Closing-the-loop in nvisionip: integrating discovery and search in security visualizations. In: Proceedings of the Visualization for Computer Security. (VizSEC 05), Minneapolis, USA, 2005. 75-82

[11] Taylor T, Brook S, Mchugh J. Netbytes viewer: An entity-based netflow visualization utility for identifying intrusive behavior. In: Proceedings of the 4th International Workshop on Visualization for Cyber Security, Sacramento, USA, 2008. 101-114

[12] Fischer F, Mansmann F, Keim D A, et al. Large-scale

network monitoring for visual analysis of attacks. In: Proceedings of the 5th International Workshop on Visualization for Computer Security, Cambridge, USA, 2008. 111-118

[13] Boschetti A, Muelder C, Salgarelli L, et al. TVi: a visual querying system for network monitoring and anomaly detection. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security, Pittsburgh, USA, 2011. 1-10

[14] Braun L, Volke M, Schlamp J, et al. Flow-inspector: a framework for visualizing network flow data using current web technologies. *Computing*, 2014, 96(1): 15-26

[15] Zhou F, Shi R, Zhao Y. NetSecRadar: A Visualization System for Network Security Situational Awareness. *Cyberspace Safety and Security*, 2013, 8300: 403-416

[16] Zhao Y, Zhou F, Fan X, et al. IDSRadar: a real-time visualization framework for IDS alerts. *Science China Information Sciences*, 2013, 56(8): 1-12

[17] 赵建秀, 王洪国, 邵增珍等. 一种基于信息熵的时间序列分段线性表示方法. 计算机应用研究,2013, 30(8): 2391-2394

[18] 夏秦, 王志文, 卢柯. 入侵检测系统利用信息熵检测网络攻击的方法. 西安交通大学学报, 2013, 47(2): 14-20

[19] Shiravi H, Shiravi A, Ghorbani A A. A Survey of Visualization Systems for Network Security. *IEEE Transactions on Visualization and Computer Graphics*, 2011, 1(1): 1-19

Research on applying information entropy time series and TreeMap to NetFlow visualization

Zhang Sheng[*,**], Shi Ronghua[*], Zhao Ying[*], Zhou Fangfang[*]

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

([**]Modern Educational Technology Center, Hunan University of Commerce, Changsha 410205)

Abstract

Considering that the management and analysis of the NetFlow log are becoming more difficult because of the NetFlow log's increase in size and changing speed, a Visualization system for analysis of the NetFlow log by using the Time series map combined with the TreeMap according to the concept of network security visualization, was constructed to quickly, effectively identify network attacks and abnormal events in networks. By focusing on the six characteristics of information entropy, the system can successfully oversee the network security situation against the Time Series. At the same time, it can drill down into the details of invasion by using the TreeMap. The system also uses an image feature rule to construct visual figures for attack analysis and pattern exploration. Through the analysis of the VAST Challenge2013 competition data on this system, it was showed that the system can intuitively capture the network security status from the macro and micro levels, as well as effectively identify network attacks and give the support in decision-making.

Key words: network security visualization, security situational awareness, NetFlow, information entropy, Time Series, TreeMap