基于信息熵的网络流异常监测和三维可视方法

陈 鹏,司 健,于子桓,王蔚旻
CHEN Peng, SI Jian, YU Zihuan, WANG Weimin

中国电子科技集团公司 第二十八研究所 第一研究部,南京 210007
The First Research Department, No.28 Research Institute, China Electronics Technology Group Corporation, Nanjing 210007, China

CHEN Peng, SI Jian, YU Zihuan, et al. Flow abnormity supervision based on information entropy and 3D visualization. Computer Engineering and Applications, 2015, 51(12):88-93.

Abstract: Through the analysis of network traffic, the network condition reflecting, abnormal behavior mining, network security situation awareness are enabled. Large scale network flow has mass data and wide range dimensions. Aiming at these features, in order to monitor network running situation and abnormity and improve the users' awareness experience, this paper puts forward a kind of quasi real time flow reporting mechanism, designs a flow monitoring system based on 3D visualization, and combines with the flow abnormity mining method based on information entropy, through manual monitor and data mining, realizes abnormal flow visualization monitoring. It presents the monitoring system design scheme and implementation results, resolves the hard problem of network flow visualization, puts forward a kind of traffic situation scheme which is more intuitive, improves the users' network situation awareness capability.

Key words: netflow; traffic collection; information entropy; abnormal flow; flow visualization; traffic monitor system

摘 要: 通过分析网络流量可以反映网络运行情况,挖掘异常行为,感知网络安全态势。为了监测网络运行状况和流量异常情况,提高用户对网络流量态势的感知体验,针对大规模网络流量的数据量大和维度广的特点,提出了一种准实时流量数据报出机制,设计了基于三维可视化的流量监测系统,并结合基于信息熵的流量异常挖掘方法,通过人工监测和数据挖掘,实现了异常流量可视化监测,提高了异常检测成功率。给出了监测系统的设计方案和实现结果,解决了网络数据流从抽象到具象的可视化问题,提供了一种更加直观的态势展现方案,提高了用户对网络态势的感知认识能力。

关键词: 网络流; 流量采集; 信息熵; 异常流量; 流量可视化; 流量监测系统
文献标志码: A **中图分类号:** TP393 doi:10.3778/j.issn.1002-8331.1408-0111

1 引言

长久以来,网络安全就是人们关注的热点问题。尤其近年来,产生了很多新型网络攻击手段,互联网上的黑客攻击事件呈指数增长,给网络安全防护带来了严峻挑战。然而,在被监控网络层次复杂、规模庞大、信息服务种类多样的情况下,安全防护系统会从各种安防设备采集到大量数据,与此相矛盾的是,安防系统的分析资源和分析能力是有限的,在海量信息的冲击下,安防系统很容易过载,从而造成安全事件的漏报和误报。

另外,现在异常发现方面的人工智能水平还无法达到完全模仿人类的程度,因此,在安全防护系统中引入人工辅助是可能而且必要的。发生网络攻击、网络病毒、网络故障、突发访问和新用户加入时,网络流量会体现出不同于正常模式的表征,因此,通过网络流量可以反映当前网络的安全态势情况。而人可以分辨图像在位移、色彩、形状、模式等方面的细微变化[1],通过对正常网络态势可视化图像的观察和学习,建立正常模式的认知[2],当异常网络行为发生时,安全管理员可以马上感知,并借助机器准确定位发生问题的区段和性质,从而降低漏报和误报概率。

作者简介: 陈鹏(1978—),男,高级工程师,研究领域为计算机网络;司健(1985—),男,助理工程师,研究领域为网络安全;于子桓(1990—),男,助理工程师,研究领域为计算机网络;王蔚旻(1983—),男,高级工程师,研究领域为计算机网络。E-mail:sijian2012@sina.com

收稿日期: 2014-08-25 **修回日期:** 2014-10-24 **文章编号:** 1002-8331(2015)12-0088-06
CNKI **网络优先出版:** 2015-02-11, http://www.cnki.net/kcms/detail/11.2127.TP.20150211.1039.024.html

一个普通显示屏的图片信息量大概是150 MB,而同样面积的文本的信息量却只有100 KB[2],文本信息密度相对而言比较有限。网络管理员不可能通过整屏幕的文字来感知分析海量网络数据,可视化技术已经成为网络态势感知的一个重要组成部分。

相对于机器对数据的认知和处理方式,人更适合于通过画面的方式来认识世界,研究数据表明,人类可以通过可视化的方式以150 Mb/s的速度处理信息[2],可见,相对于枯燥乏味的数据,图像更容易被人所理解和接受。因此,提供一种更友好的人机交互手段,对于用户更好地理解和认识网络流量、发现异常是很有必要的。而MRTG(Multi Router Traffic Grapher)[3]、NVisionIP[4]、VisFlowConnect-IP[5]等一些现有的流量可视化系统流量实时性较差,且不支持对异常流量的自主挖掘。本文介绍了一种网络流监测系统,通过人在回路(man in loop)[2]的方式,直观展现了抽象的网络流量,并结合基于信息熵的异常挖掘算法,实现对网络流量的可视化监测和异常分析。

2　网络流生成原理

网络流定义了一条在源和目的之间的单方向数据流,具体来说,一个网络流至少包含了源IP地址、目的IP地址、源端口号、目的端口号、三层协议类型和TOS字段。

现有的网络流量生成方法包括NETFLOW、sFlow和SNMP。NETFLOW通过监测流量的活跃和空闲状态,进行流量数据报出[6]。sFlow是一种"永远在线"技术,sFlow由代理端和中心处理端组成,可用于超大网络环境下的流量线速分析[7]。NETFLOW和sFlow都使用采样技术,常用的流采样包括Poisson采样、周期采样和自适应采样等。

Poisson采样的采样时间间隔符合参数为λ的指数分布,新样本的产生不可预测,泊松过程是无偏的。周期采样包括基于时间和基于报文两种情况,基于时间就是根据固定时间间隔采样,基于报文就是根据固定的报文数量采样。自适应采样基于神经网络、小波分析等预测算法预测数据变化幅度,动态调整采样间隔,流量抖动剧烈的时段缩短采样周期以反映流量变化,流量抖动平缓的时段加大采样周期以节省系统带宽[7]。

3　流量采集处理方法

3.1　流量采集报出

将流量采集点网卡设置为"混杂模式",接收到达本网卡的所有镜像流量。为了精确分析,不对数据流做抽样处理。捕获到流数据后,采集点将该流加入缓存,对于UDP连接,如果流在3 s内都是空闲无数据的,认为流结束,将流数据从缓存清除;对于TCP连接,连接完成(FIN)或者被重置(RST)时,认为流结束,将流数据从缓存清除。

设定流量处理中心IP地址,流数据按照1次/s的频率向这个IP地址周期性报出由源IP、目的IP、源端口、目的端口和协议类型构成的五元组流量数据,每次报出时更新流量大小和包数,向流量处理中心准实时反映采集节点的流量情况。流的报出示意图如图1所示。

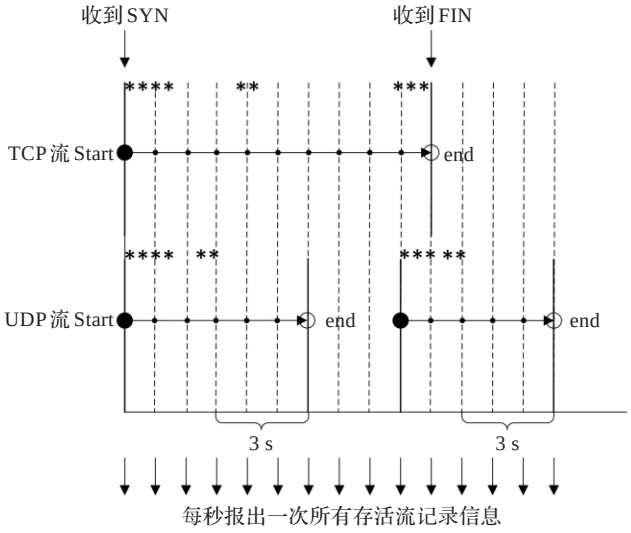


图1　网络流报出机制

3.2　流量集中处理

为了支持大规模网络的流量态势可视化监测,本文提出了一种基于网络流量镜像的数据包采集分析方案,如图2所示,用普通网卡镜像采集交换机上内外网之间端口流量,流量处理中心对各个网段采集点的流量数据进行归并处理,最后生成流量态势。本方法数据流实时性强,可以体现不同网段之间的流量关系,成本低,灵活性强。

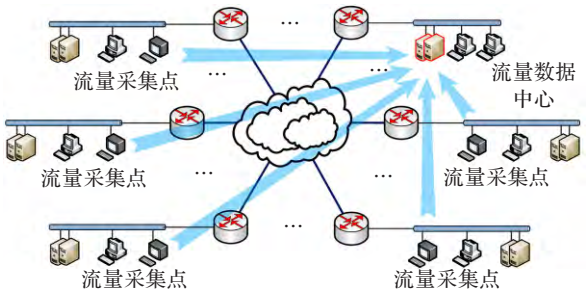


图2　流量整合处理

流量处理中心支持流量数据的实时分析和长时间跨度历史数据的统计查询,处理流程如图3所示。

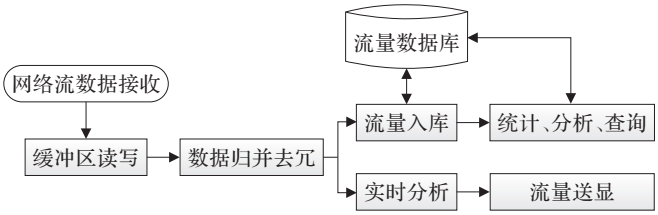


图3　流量处理流程

在大规模网络中,流量即使通过流格式的缩减,但是数据包数量不会缩减。在流量处理中心用数据缓冲

区来存储已接收,但暂时没有被处理的流量数据,并使用互斥锁保证缓冲区的数据一致性。

当两个采集点网段之间产生流量交互的时候,处理中心会收到来自不同采集点的相同网络流数据。流量处理中心将具有相同五元组(源IP、目的IP、源端口、目的端口和传输协议)的网络流数据归并为同一个流,并以五元组为查询键值,将网络流数据存储在哈希表中。

另外,对于UDP连接,每一个哈希流节点都有一个保活时间,当相同五元组的网络流数据到达后对应的哈希流节点保活时间更新,定时查询哈希节点保活时间,认为超时节点的流已经断开,删除超时节点。对于TCP连接,如果收到FIN报文,则认为TCP断链,删除相应的哈希节点。

流量处理中心一方面对流量数据进行实时的信息熵计算,分析异常流量,并上图显示;另一方面将流量数据入库,从而支持历史数据的查询统计,以及大跨度的流量分析。

4 流量异常监测

4.1 基于熵的异常挖掘

熵体现了系统的不确定性和无序性,一个系统越有规律,熵值就越小;越无序混乱,熵值就越大。1984年香农把熵的概念引入信息论中,并基于概率统计理论给出了信源信息熵的定义:

$$H=-\sum_{i=1}^{N}P_i\,\mathrm{lb}\,P_i \tag{1}$$

信源包含 N 个以一定概率出现的事件或者信息基元,其中 H 表示信源信息熵,P_i 表示某个事件发生的概率。

网络流量包含有多种特征属性,比如源IP、目的IP、源端口、目的端口等。这些特征属性的取值在一定时间间隔的网络流量中具有一定的分布特性,而熵值可以有效体现数据统计分布的集中和松散程度,当网络流量发生异常时,对应的网络信息熵会产生相应变化,因此可以通过计算网络流量数据的信息熵,分析网络流量分布特征。

令待分析的 m 个网络流特征属性为 $A_i(i=1,2,\cdots,m)$,n_{ij} 表示属性 i 的第 j 个取值在一个统计时间间隔中的流个数,其中 $j=1,2,\cdots,y_i$,y_i 表示属性 i 的取值种类个数,并令 $S_i=\sum_{j=1}^{y_i}n_{ij}$ 表示属性 i 的总的流个数,则属性 i 的信息熵定义为:

$$H(A_i)=-\sum_{j=1}^{y_i}\frac{n_{ij}}{S_i}\mathrm{lb}\frac{n_{ij}}{S_i} \tag{2}$$

在网络流量分析中需要对多个流量属性进行分析,分别定义 $H(srcIP)$、$H(dstIP)$、$H(srcPort)$、$H(dstPort)$ 为源IP熵、目的IP熵、源端口熵和目的端口熵,分析流量分布的集中、分散情况,统计流量分布特征。网络流量会随着时间推移呈现不同的分布状态,把时间轴划分为等长的时间间隔,在时间间隔内进行属性熵值的计算和预测,然后根据观察熵和预测熵的偏差来判断是否发生异常。

预测下一时间间隔信息熵使用指数平滑法,指数平滑法又叫指数加权平均法,通过对历史数据分配按照指数递减的权值,加权平均得到预测值,该方法可以有效消除历史数据中的随机抖动,找出主要发展趋势。指数平滑计算公式为:

$$S_f(t)=\begin{cases}\alpha\cdot S_0(t-1)+(1-\alpha)\cdot S_f(t-1), & t>2\\ S_0(1), & t=2\end{cases} \tag{3}$$

其中 $S_f(t)$ 表示 t 时段的预测熵,$S_0(t-1)$ 表示 $t-1$ 时段的观察熵,α 为平滑因子,平滑因子是根据观察熵所占的权重来确定。

$$\alpha=1-\exp\left(\frac{\ln(1-w)}{n}\right) \tag{4}$$

其中 w 表示观测熵所占的百分比权重,n 表示所取历史数据点的数目。

采用指数平滑法根据短期内历史数据进行信息熵的预测,令 $\hat{h}_{t,i}$ 为属性 i 的 t 时段预测熵,$h_{t-1,i}$ 为属性 i 的 $t-1$ 时段观察熵,$\hat{h}_{t-1,i}$ 为属性 i 的 $t-1$ 时段预测熵,指数平滑法计算公式为:

$$\hat{h}_{t,i}=ah_{t-1,i}+(1-a)\hat{h}_{t-1,i} \tag{5}$$

将最近 k 次的预测熵 $\hat{h}_{t-1,i},\hat{h}_{t-2,i},\cdots,\hat{h}_{t-k,i}$ 递推代入上式,可得:

$$\begin{aligned}\hat{h}_{t,i}=&ah_{t-1,i}+a(1-a)h_{t-2,i}+\cdots+\\&a(1-a)^{k-1}h_{t-k-1,i}+a(1-a)^k h_{t-k,i}\end{aligned} \tag{6}$$

可见预测熵为历次观测熵的指数平均加权的结果,而且对历史观测熵的赋权会随着时间的增加而呈指数递减。

根据概率统计理论,信息熵的正常区间范围是 $[\hat{h}_{t,i}-3\sigma_{t,i},\hat{h}_{t,i}+3\sigma_{t,i}]$,其中 $\sigma_{t,i}$ 表示时段 t 内,属性 i 信息熵的标准差。

$$\sigma_{t,i}=\sqrt{\frac{1}{k-1}\sum_{j=t-k}^{t-1}(h_{ij}-\hat{h}_{ij})^2} \tag{7}$$

当观察熵和预测熵的偏差超过 $3\sigma_{t,i}$,即 $|\hat{h}_{t,i}-h_{t,i}|\geqslant 3\sigma_{t,i}$ 的时候[8],认为网络流量发生异常,触发告警。

4.2 异常流量的熵模式

以DDOS攻击为例,如果多台主机对一台服务器发起攻击,网络中会产生很多源IP不同,目的IP相同的流,这些攻击流量提高了网络流总的源IP的离散程度和目的IP的集中程度,从而导致源IP熵增大和目的IP熵减小。

和DDOS攻击类似,端口扫描、主机扫描、大流量数据传输等网络行为都会导致网络流量异常,每一种异常行为反映到网络信息熵变化趋势上都体现出不同的对应模式。常见典型模式如表1所示。

表1　异常流量熵模式

异常名称	说明	$H(srcIP)$	$H(srcPort)$	$H(dstIP)$	$H(dstPort)$
端口扫描	对单个主机或者服务器端口开放情况进行扫描			减小	增大
主机扫描	单个主机对多个主机进行开机探测			增大	减小
DDOS攻击	拒绝服务攻击	增大		减小	
异常大流量	点到点之间流量忽然增大	减小		减小	
蠕虫扩散	通过自主复制在网络中扩散病毒	增大		增大	减小

通过监测网络流的源IP、目的IP、源端口和目的端口的信息熵,对发生熵值异常的属性进行异常模式匹配,判定发生的异常类型。然后根据异常类型在发生异常的时段内,遍历网络流,寻找异常发起源。如果发现异常规则库中不存在的模式类型,则记录熵值变化特征,评估异常影响,录入新型异常规则。

5　系统设计

网络流监测系统的主体是一个三维立方体(下面简称为体视图),立方体中的点由源IP、目的IP和目的端口三维坐标来描述,X轴、Y轴和Z轴分别代表源IP、目的IP和目的端口。支持全IP地址空间流量显示,和任意长度掩码网段归并。

当一个连接的数据传输开始的时候点显示,数据传输停止的时候点消失,在这里把一个光点称为一个连接,通过点来展示网域空间节点连接关系、数据流向和服务类型。本系统的主要功能包括以下几个方面。

5.1　可视化设计

5.1.1　三维视图拆解

三维立方体的默认范围是全域地址,包含了A、B、C类地址,立方体的X轴和Y轴上包含了232个点,地址空间很大。当连接多到一定程度的时候,点和点之间会相互交叠,影响观察。通过立方体定位,可以在三维空间中定位到感兴趣的网络区段,然后使用拆解功能,将感兴趣的网络区段拆分放大,方便用户查看。三维视图拆解如图4所示。

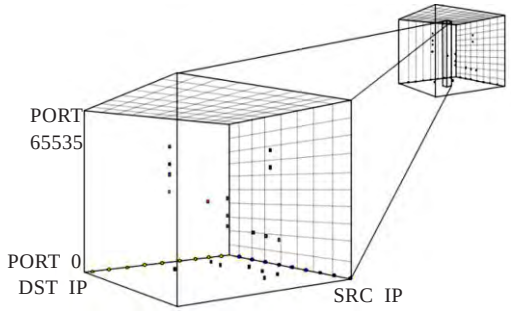


图4　三维视图拆解

5.1.2　二维视图切换

二维视图包括面视图和点视图,面视图是体视图的切面,固定三维立方体$X(Y)$坐标,截取$YOZ(XOZ)$平面,此平面视图上的点表示某一源(目的)IP和多个目的(源)IP建立的连接;固定Z坐标,获取XOY平面,此视图上的点表示某一目的端口上的连接情况。面视图的典型应用包括查看某个服务器给多个节点提供服务的流量态势,查看某个目的端口上的连接情况等。面视图如图5所示。

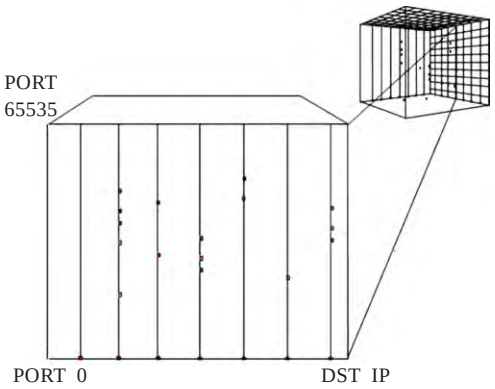


图5　面视图

5.1.3　点视图切换

点视图是体视图的另一种观察方式,点视图由四条平行线组成,这四条线分别代表源IP、目的IP、目的端口和应用报文类型。点视图中一条完整的折线表示一个连接,图中点的半径大小反映了对应连接的流量大小。点视图是体视图的简化形式,目的在于以更简洁的方式体现网络连接。点视图如图6所示。

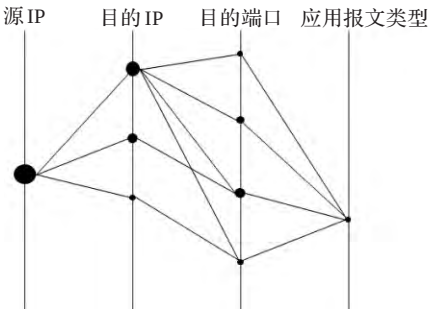


图6　点视图

5.2　系统框架

如图7所示,网络流监测系统由cubeview组件、界面框架、流量数据管理、异常监测和流量数据采集四个部分组成[9]。流量数据采集部分部署在各个局域网的流量采集点,cubeview组件、界面框架和流量数据管理部署在流量处理中心。采集点生成流量数据后上报处理中心,处理中心在流量数据管理组件和界面框架的支持下,通过cubeview上图显示,并监测异常流量[10]。

流量数据采集模块采集原始流量,按照第2章的流量生成机制,向流量处理中心报出所在局域网的流量信

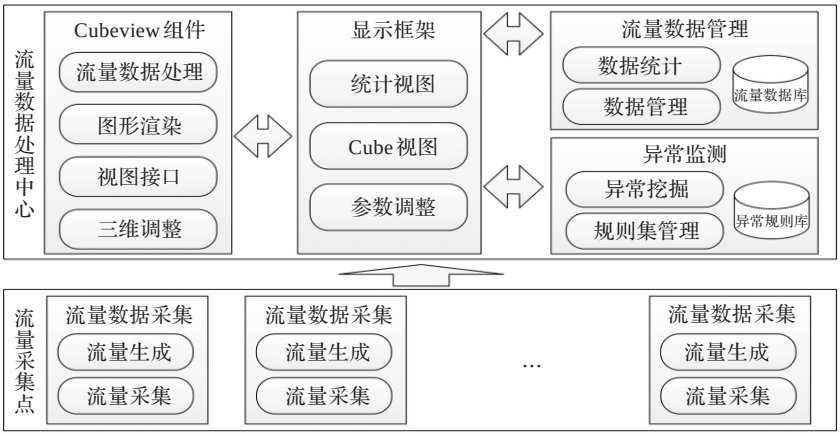


图7　系统框架

息[11]。显示框架收到原始流量信息后,调用cubeview组件中提供的绘图接口,进行二维和三维图形绘制,并支持三维视角调整和二三维视图切换。

同时,流量数据管理模块使用双端队列管理最近的流量数据,支持异常流量实时监测和历史态势重演,并为历史数据查询提供流量数据库操作接口。另外,异常监测模块通过基于信息熵的短期监测和历史数据的长期挖掘,结合异常规则集,探测异常流量,支持异常规则集的手动输入和机器自主学习。

流量监测界面包括参数区、视图区、统计区三部分,参数区位于系统界面左侧,用于体视图拆解、面视图截取、二三维切换等参数配置;视图区位于系统界面中间,是监测系统的主体,用于展示体视图、面视图和点视图;统计区位于系统界面右侧,用于展示系统的实时连接统计曲线和柱状图。流量处理中心系统实现截图如图8所示。

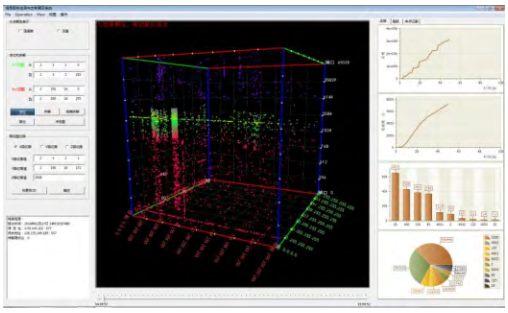


图8　监视界面

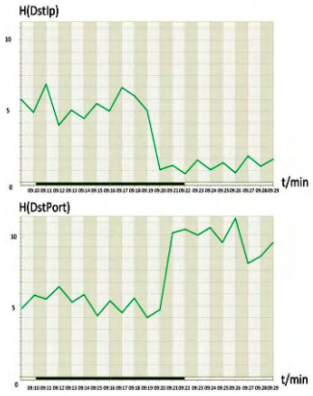
6　结果分析

当出现异常流量的时候,不仅网络流的信息熵值曲线会发生相应变化[12],而且视图区的点分布会根据特定异常体现出不同的模式[13]。实验中通过模拟异常网络行为,验证系统的异常发现效果。

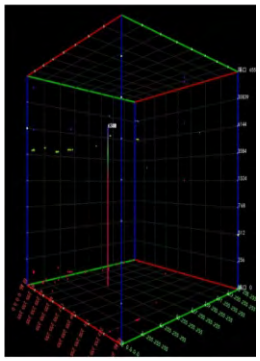
6.1　端口扫描实验

端口扫描是黑客攻击前进行网络嗅探的关键步骤,在实验网络环境中的主机(102.14.10.10)上部署端口扫描软件,对特定服务器(97.20.20.101)进行端口扫描,扫描发起端和被扫描端IP都是单一值,扫描端口呈平均分布。

目的IP熵和目的端口熵的系统监测曲线如图9(a)所示,可以看出,在9:20时段,目的IP熵急剧下降,目的端口熵急剧上升,这个时段正是端口扫描发生的时间,网络中会产生很多目的IP相同,目的端口呈平均分布的流,这些异常流量提高了网络流总的目的IP的集中程度和目的端口的离散程度,从而导致目的IP熵减小和目的端口熵增大。



(a)目的IP和目的端口熵值曲线



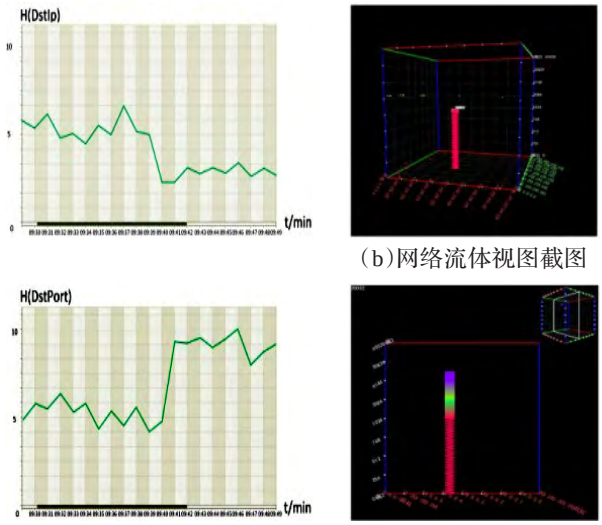
(b)网络流体视图截图

图9　端口扫描实验截图(单目的主机)

机器挖掘自动告警的同时,网络流可视化监测界面同步展示端口扫描异常流量的分布态势,根据体视图展示原理,端口扫描呈现的效果是垂直于 *XOY* 平面的一条直线,实验截图如图9(b)所示。

配置端口扫描软件同时对多个主机进行扫描,扫描发起IP为单一值,被扫描端IP为若干个不同的值,扫描端口呈平均分布。目的IP熵和目的端口熵的系统监测曲线如图10(a)所示,可以看出,熵值变化的模式和单机端口扫描基本一致,只是目的IP熵下降没有单机端口扫描的情况那么剧烈,这是因为被扫描的多个主机具有多个IP地址,相对的目的IP集中程度弱一些。同时,从网络流可视化效果呈现为垂直于 *XOY* 平面的折线,实验截图如图10(b)所示。

为了便于查看,固定发起扫描的源IP,切换到二维面视图,系统截图如图10(c)所示。该视图展现了102.14.10.10主机上向多个目的主机进行端口扫描的态势,面视图呈



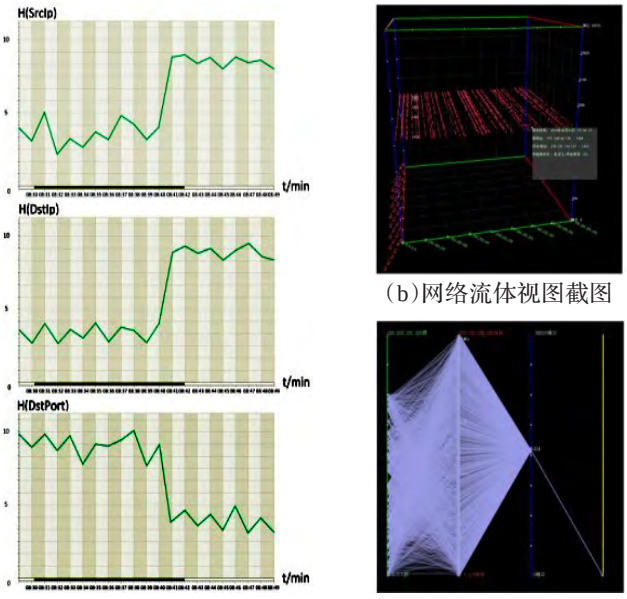
(a)目的IP和目的端口熵值曲线　(b)网络流体视图截图　(c)网络流面视图截图

图10　端口扫描实验截图(多目的主机)

现折线状。

6.2　蠕虫扩散实验

感染蠕虫病毒的主机会传播蠕虫拷贝到网络上其他主机,其他主机进一步拷贝扩散蠕虫病毒,因此,如果不受控制,蠕虫病毒会在网络中呈爆炸型扩散。蠕虫扩散一般使用单一端口(一般是1434),在实验环境中部署蠕虫模拟器,源IP熵、目的IP熵和目的端口熵的系统监测曲线如图11(a)所示。



(a)源IP、目的IP和目的端口熵值曲线　(b)网络流体视图截图　(c)网络流点视图截图

图11　蠕虫扩散实验截图

网络流可视化监测界面同步展示蠕虫扩散异常流量的分布态势,根据体视图展示原理,端口扫描呈现的效果是同处于1434端口的平面的一系列平行线,实验截图如图11(b)所示。蠕虫扩散对应的流量点视图如图11(c)所示,从点视图可以看出多个被感染主机向其他主机扩散蠕虫病毒的异常流量态势。

7　结束语

可视化人机界面是监测系统和用户交互的关键平台,在人因越来越重要的今天,可视化技术已经成为网络态势感知的关键部分。本文设计的针对大规模网络的网络流生成和可视化监测系统,可以较好地为用户展示网络流量态势,发现异常流量。另外,网络环境日趋庞杂和网络流量的日渐增长,流量处理中心需要对多点报出的流量数据进一步提高处理和管理效率[14],针对流量异常告警,进一步提高发现异常流的能力,丰富可视化展现元素[15]。

参考文献:

[1] Heberlein L T,Dias G V,Levitt K N,et al.A network security monitor[C]//IEEE Computer Society Symposium on Research in Security and Privacy,1990:296-304.

[2] Bearavolur,Lakkarajuk,Yurcik W.NVisionIP:an animated state analysis tool for visualizing netflows[C]//Proc of FLOCON Network Flow Analysis Workshop,2005.

[3] Lakkarajuk,Yurcikw,Lee A J.NVisionIP:NetFlow visualizations of system state for security situational awareness[C]//Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security,Washington,DC,2004:65-72.

[4] Yurcik W.Visualizing NetFlows for security at line speed:the SIFT tool suite[C]//Nineteenth Systems Administration Conference,2005.

[5] Yin Xiaoxin,Yurcik W,Slagell A.The design of visFlowconnect-IP alink analysis system for ip security situational awareness[C]//Third IEEE International Workshop on Information Assurance,2005:141-153.

[6] 黄晓琳.基于MRTG的校园网网络流量监控技术研究[J].科技创新导报,2013(3).

[7] 严承华,程晋,樊攀星.基于信息熵的网络流量信息结构特征研究[J].信息网络安全,2014(3).

[8] 第文军,薛丽军,蒋士奇.运用网络流量自相似分析的网络流量异常检测[J].兵工自动化,2003,23(6).

[9] 何长林,党小超.网络测量及流量采集技术综述[J].计算机时代,2011(7):14-15.

[10] 范亚国.基于sFlow的网络链路流量采集与分析[D].武汉:武汉理工大学,2008.

[11] 李达.基于流量采集和参数测量的网络可靠性分析系统的设计与实现[D].长沙:国防科技大学,2009.

[12] 苏锡鑫,苏伟,刘颖.基于熵的流量分析和异常检测技术研究与实现[J].计算机技术与发展,2013,23(5):120-123.

[13] Denning D E.An intrusion detection mode[J].IEEE Transactions on Software Engineering,1987,13(2):222-232.

[14] 周光霞,孙欣.赛博空间对抗[J].指挥信息系统与技术,2012,3(2):6-10.

[15] 芮平亮,王芳.网络电磁空间防御作战能力需求分析[J].指挥信息系统与技术,2011,2(1):1-5.