计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(4):1−8 [doi: 10.15888/j.cnki.csa.006836]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Netflow 的网络安全大数据可视化分析①

王全民, 韩晓芳

(北京工业大学 信息学部, 北京 100124)
通讯作者: 韩晓芳, E-mail: han_xiao_fanglv@163.com

摘　要: 近年来网络安全日志数据呈现出爆炸式的增长, 但现有的可视化技术难以支持高维度、多粒度的Netflow 日志实现完善的可视化分析. 因此本文提出了一种全新的网络安全可视化框架设计方案, 采用三维柱状图展示 Netflow 日志的流量时序图, 以帮助用户快速了解和掌握网络中的异常时刻. 引用信息熵算法针对平行坐标轴的维度数据进行处理, 便于用户对多维度图形的理解, 利用矩阵图、气泡图和流量时序图进行细节分析, 最后利用该系统实现了对 DDOS 攻击和端口扫描攻击的网络异常案例分析. 研究证明本系统丰富的可视化图形以及简单易用的协同交互, 能较好的支撑网络安全人员从网络整体运行状态分析, 到定位异常时刻、监测网络行为细节的全部过程.

关键词: 网络安全可视化; 平行坐标图; 矩阵图; 三维柱状图; DDOS 攻击; 端口扫描攻击

引用格式: 王全民,韩晓芳.基于 Netflow 的网络安全大数据可视化分析.计算机系统应用,2019,28(4):1–8. http://www.c-s-a.org.cn/1003-3254/6836.html

Cyber Security Visualization Analysis Based on Netflow

WANG Quan-Min, HAN Xiao-Fang

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: In recent years, the network security log data shows explosive growth. However, the existing visualization technology is hardly to support the analysis of high dimensional and multi-granularity NetFlow log data. In order to make advantages of the visualization technology, this study proposes a new network security visualization framework to paint the picture, uses the three dimensional histogram which help users to quickly master the abnormal moment of network shown by the Netflow log data, uses information entropy algorithm to process multi-dimensional data, makes use of the matrix chart, bubble diagram, and line chart to synthesize analyzed data in detail. Finally, we carry out an experiment to test the process of DDOS attack, and Port Scanning Attack can be detected easily by proposed system. The study proves that the system which has rich visual graphics and provides simple collaborative interaction can better help network security personnel to analyze the entire network behavior process.

Key words: network security visualization; parallel coordinate plots; matrix diagram; 3D bar; DDOS; port scanning attack

1　引言

Netflow 日志可以提供非常精准的流量测量, 被广泛应用在 DDOS 监控、入侵检测、流量统计等方面, 所以对 Netflow 数据的研究具有极强的实用意义, 很多研究者将此可视化研究的切入点. 如 Nunnally T[1]将传统的平行坐标轴进行了拓展, 提出采用 3D Parallel Coordinate 方法对安全数据进行可视化, 并取得了较好的效果. 但该方法在数据量过大时, 线条之间会发生重

① 收稿时间: 2018-10-06; 修改时间: 2018-10-23; 采用时间: 2018-11-05; csa 在线出版时间: 2019-03-28

叠, 难以从中发现有用的信息. Flow-In-Spector[2]采用多图综合的可视化技术实时显示网络流数据, 实现了直方图、力引导图、辐射图的绘制. 吴亚东等[3]为了增强分析系统的可交互性, 设计了一种三维多层球面空间的可视化模型. 陈鹏等[4]利用信息熵的流量异常数据挖掘算法, 提高了流量异常检测的成功率, 并实现了一个三维可视化的流量监测系统. 张胜等[5]针对 Netflow 日志用树图和时间序列图结合的方式实现了可视化系统, 但当数据量较大时, 二维的时间序列图会造成图形覆盖, 且树图会占用较大的屏幕空间, 对量大、维度多的数据展示很不友好.

在可视化分析方向, 多视图的关联分析和可交互式查询对帮助网络安全管理人员从多维度观测当前网络状态至关重要. 比如赵颖等[6]通过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、相似度扩展树视图分析网络流量日志数据. ENTVis[7]利用雷达图、矩阵图等可视化方法支持基于信息熵的网络攻击流量特征分析, Shi 等[8]利用改进雷达图进行网络事件关联分析. 以上技术虽然通过特征抽取、降维、采样或聚合的方法减少了数据项和数据维度, 但交互性并不好.

综上所述, 针对 Netflow 日志的可视化分析技术研究已有很大突破, 但在多视图综合交互以及多维度数据可视化方面仍有很高的提升空间. 因此, 本文重点提出了用三维柱状图代替二维坐标图展示数据, 它可以提供旋转、缩放、筛选等多种交互方式解决密集数据遮掩的问题, 针对平行坐标轴引入信息熵算法实现多维度特征的统一化处理, 能较好的展示多维度数据. 在最后的在案例分析中, 利用本文的可视化分析系统实现了端口扫描、DDos 攻击网络异常行为的分析.

2 系统可视化框架

为了帮助网络安全人员从整体到局部, 循序渐进的对整个网络态势有准确的掌控, 本文设计了一个两层的可视化框架, 如图 1、图 2 所示. 图 1 中的三维柱状图帮助用户了解当前网络态势, 平行坐标图对应网络整体流量时序的多维特征分析, 用以分析各个维度的详细信息. 图 2 细节图提供了 3 个可视化视图, 多维气泡图和矩阵图详细介绍了多种网络活动主体在某个时间段内的细节分布, 流量时序视图对应某个网络活动主体的时序特征分析.

2.1 三维柱状图

任何网络行为的发生都会在 Netflow 日志中留下一条 flow 记录, 该记录中的字节数反映了此次网络行为占用的流量大小. 当有可疑的网络行为发生时, 可以通过分析 Netflow 日志中的字节数, 判断其行为是否异常, 比如当有大量的攻击行为产生时, 网络流量的数量级别会显著升高.

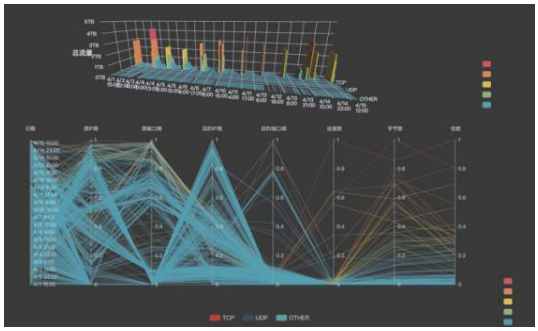

图 1 概览图

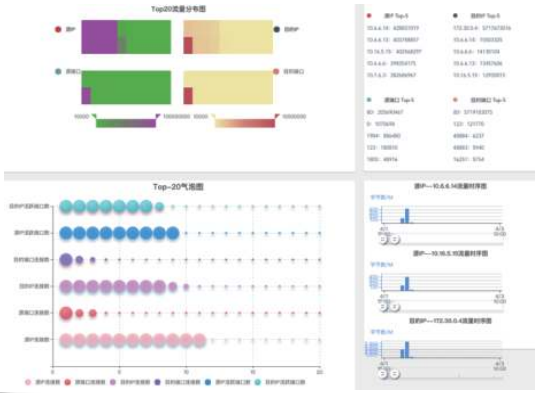

图 2 细节图

以前针对流量的展示方法大多是二维的折线图和柱状图, 无法从多个角度分析当前流量的分布, 本文的三维柱状图 (如图 3) 按时间序列统计一段时间内不同协议下的字节数, 从图 3 柱状图的高低和颜色的深浅, 用户可以快速判断出 TCP 协议下, 颜色较深且柱状图较高的时间段内, 流量异常增多, 而在整个时间段内 UDP 协议和其他协议下的流量较少. 用户可点击图 3 右下角的方块 (从下到上, 流量大小按区间递增) 快速筛选目标数据, 也可利用放大、平移和旋转功能, 快速查看被遮挡区间内的详细数据. 算法 1 给出了本文采用 Echarts 实现三维柱状图的核心算法.

算法 1. 三维柱状图核心算法

```
grid3D: {
  boxWidth: 400, boxDepth: 80,
  left: -100, top: 0,
  viewControl: {distance: 320},
  light: {
    main: {intensity: 1.2, shadow: true}, ambient: {intensity: 0.3}
  },
  axisPointer: …
  },
  dataset: {// 三维数组模型
    dimensions: ['days', 'protocolCode', 'srcAllBytes'],
    source: data
      },
series: [// 三维坐标系数据配置
  {type: 'bar3D', name: '…',
  symbolSize: symbolSize,
  shading: 'lambert',
  encode: {// 匹配 dataset 数据
  x: 'days', y: 'protocolCode',
  z: 'srcAllBytes',
  tooltip: [0, 1, 2], label: 2
  },
……
}]
```

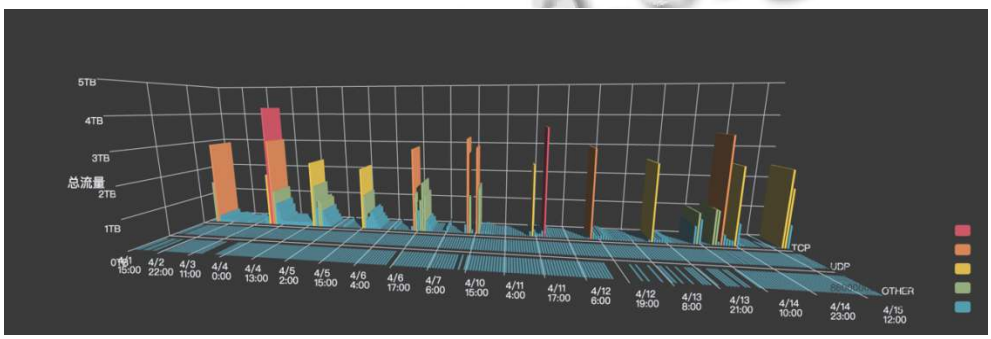

图 3 三维柱状图

2.2 平行坐标图

分析人员通过观测三维柱状图中流量的大小, 能快速定位异常时刻. 为了判断该时刻下可能发生的异常行为类型, 分析人员需要同时观测 Netflow 的多个维度特征, 以做出更准确的决策.

Netflow 数据中的源地址 (srcIP)、源端口 (srcPort)、目的地址 (destIP)、目的端口 (destPort)、连接数 (link)、字节数 (destTotalBytes) 和包数 (destPackets), 这 7 个维度最能体现网络安全态势的变化. 由于各个特征间的数量级别差异较大, 且不同的维度统计分析的指标不同, 无法直接使用以上 7 个特征的统计值作为分析点. 信息熵能有效度量网络活动的随机特征, 它是一种对异常分布很敏感的度量参数, 体现了数据分布的不确定性和无序性, 数据分布越有规律, 熵值越小; 越无序混乱, 熵值越大. 下面我们以目的端口熵为例, 介绍网络信息熵的计算方法, 设某一时间段内, 流量端口号集合为随机变量 Y, 则 Y 的取值空间为$Y=\{y_i, j=1,2,\cdots,N\}$, y_j为某个端口号在该时间段下出现的次数, j 实际取值范围 0–65535, S 为该时间段内总的目的端口数, 则该时间段内的目的端口熵为:

$$H(Y)=-\sum_{j=1}^{N}\frac{y_j}{S}\log\frac{y_j}{S} \tag{1}$$

例如某些类型的 DDOS 攻击是通过发送海量的攻击包到特定的目的主机, 从而达到让某个被攻击目标的服务陷入瘫痪, 这时被攻击的目的 IP 固定分布在某几个, 目的 IP 熵会变小, 攻击的源 IP 熵增大. 导致网络流量异常的行为有很多种, 常见的异常流量熵模式如表 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	对多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	
蠕虫扩散	通过自主复制在网络中扩散病毒	增大		增大	减小

平行坐标图能很好的支持用户从多个维度协同分析数据特征, 它是一种基于二维图形表达高维数据的可视化技术, 能将复杂的高维数据在平面图中展示出来, 较为节省屏幕空间, 如图 1 的平行坐标图展示了整个时间段内的数据, 利用鼠标的筛选功能, 得到某个时刻下的平行坐标图, 如图 4.

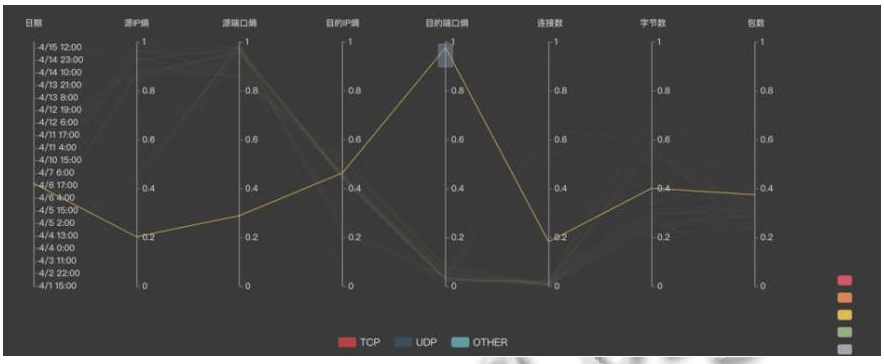

图 4 端口扫描攻击—平行坐标图

本文平行坐标图的可视化过程主要分为两步: 首选是数据处理, 如图 4 选取时间作为平行坐标轴的测量维度, 针对源 IP、目的 IP、源端口、目的端口四个维度引入信息熵算法计算, 由于各个维度的信息熵具有不同的数量级, 采用除以最大值的归一化处理, 使取值都落在 0 和 1 之间. 针对字节数、连接数和包数量, 采用统计求和的方法, 为了和信息熵的数量级保持一致, 也采用归一化处理. 其次是平行坐标图绘制, 即将数据源转换为矩阵模型, 再将矩阵映射为平行坐标系上对应的数据. 算法 2 给出了平行坐标可视化的核心算法.

算法 2. 平行坐标图核心算法

```
parallelAxis: [ //平行坐标轴的数据映射
   {dim: 0, name:schema[0].text,
   type: 'category',
   data: this.timeData, …},
   {dim: 1, name: schema[1].text},
   {dim: 2, name: schema[2].text},
   …
],
visualMap: { //视觉编码
   show: true, min: 0,
   max: 1, right: 0,
   dimension: 6,
   type: 'piecewise',
   inRange: {
      color: […]
      }
},
parallel: {
   left: '5%', right: '10%',
   bottom: 100, width: 1000,
   parallelAxisDefault: {
      type: 'value',
      silent: 'true',
   triggerEvent: 'true',
      nameLocation: 'end',
…
   }
},
Series:[ // 数据系列映射
   {name: 'TCP', type: 'parallel',
   lineStyle: lineStyle, data:this.tcpIns},
   {name: 'UDP', type: 'parallel', lineStyle: lineStyle, data: this.udpIns},
   {name: 'OTHER', type: 'parallel', lineStyle: lineStyle, data:
this.otherIns}
]
```

2.3 气泡图和矩阵图

利用平行坐标图确定攻击行为类型后, 为进一步确定网络攻击的来源、受攻击的主机和此次攻击行为对网络造成的影响, 本文选择端口和主机的各项数据特征, 包括字节数、连接数和主机的活跃端口数作为监测对象.

传统可视化方法主要利用像素图和树图, 对某时刻网络流量在主机和端口上的分布情况进行可视化, 但大量的主机和端口看起来冗余又复杂, 用户难以快速定位到感兴趣的信息. Netflow 日志主要监测系统的网络流量数据, 当某些网络攻击行为发生时, 如 DDOS、端口扫描等攻击行为会造成主机或端口号的流量激增, 此时分析流量激增的某几台主机和端口号, 就可以判断此次网络攻击行为的源, 流量较少或变化较小的主机无法为用户提供有效且丰富的决策信息, 可以忽略. 因此本文将自动过滤影响用户判断的流量较少的主机和端口数据, 仅展示各个维度 Top-20 数量级的信息,

用矩阵图结合气泡图的形式展示, 增强用户交互体验.

图 5 左侧矩阵图展示的是流量 Top-20 的主机和端口信息, 右侧展示了对应矩阵图中四个维度的 Top-5 流量信息. 如图 6 气泡图通过气泡的大小展示数量级别, 针对数量级别差异过大的情况, 本文选择缩小气泡半径大小的极差, 在不影响可视化的情况下, 展示图形中每一个级别的数据.

2.4 流量时序图

矩阵图和气泡图只展示了网络活动 IP 和端口在特定时间段内的网络态势, 对于感兴趣的主机和端口, 需要通过分析其在正常和异常时刻的网络流量变化情况, 以进一步确定他们的行为特征. 本文采用流量时序图, 如图 7, 以观测在矩阵图和气泡图中可疑的主机或端口, 此外用户可以通过时间轴上的滑块进行区间选择, 对数据集进行筛选, 如图 8.

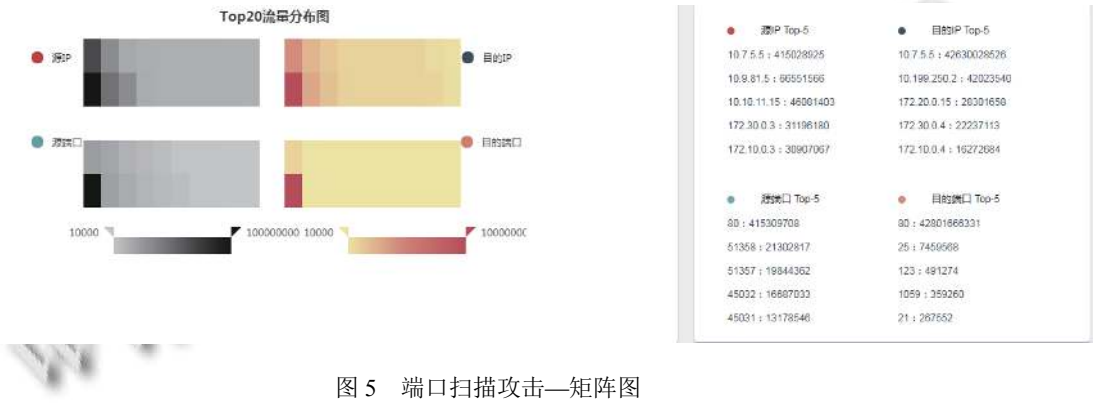

图 5 端口扫描攻击—矩阵图

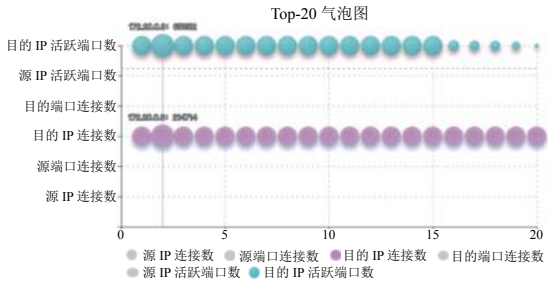

图 6 端口扫描攻击—气泡图

综上, 本系统综合多种图形实现对网络流量的可视化, 以帮助用户快速定位到网络异常时刻, 同时更便于理解图中量多且复杂的折线. 多图形间的可视化交互是本文的创新之处, 也是关联分析多维度数据的重要手段, 它凸显了可视化的意义, 提高了用户的交互体验, 也让整个系统更加紧密.

3 案例分析

本文选择可视化分析挑战赛 VAST 2013 Challenge-Mini Challenge 3[9]提供的 Netflow 日志数据作为实验数据, 图 9 是利用本文的可视化系统, 检测实验数据中潜在的网络攻击行为的分析流程.

首先对数据源进行清洗, 两周的 Netflow 数据量很大, 需要去除无效的和可视化分析无关的维度信息. 然后对可视化图形进行数据建模, 需要对平行坐标图的信息熵部分、矩阵图和气泡图中的各个端口、主机的流量、被扫描端口数等分类进行统计分析. 第二, 根据三维柱状图中方块的高度、颜色的深浅定位流量异常时刻. 第三, 获取异常时刻的平行坐标图, 综合分析各维度数据, 核对网络攻击模型, 判断网络攻击的行为. 第四, 根据判断的网络攻击行为, 分析细节图中主机、端口的流量, 端口的活跃连接数等信息, 进一步确定网络攻击行为的来源和影响的范围. 最后分析受到攻击的主机、端口及攻击源的流量时序图, 以确定发起攻击的时间范围和攻击源可能发起的其他攻击, 为用户的及时防御措施提供决策依据.

3.1 端口扫描攻击可视化分析

网络上的端口号代表该计算机提供的一种网络服务, 针对计算机上的一段端口或指定端口进行扫描, 攻击者就可以探测到该主机提供的网络服务种类, 利用这些服务的已知漏洞, 攻击者就可以开始准备攻击方法. 为了找到更多漏洞, 有的攻击者会一次性扫描六万多个端口, 此行为会造成目的端口数显著增多, 目的端口熵升高. 在平行坐标图中, 选择高亮目的端口熵趋近于 1 的时刻, 发现该时刻是 2013 年 4 月 6 日 19:00, 观测图 10 三维柱状图, 发现从 4 月 6 日 15 点开始到 4 月 6 日 20 点左右, 流量异常增多, 说明该时间段内网络受到攻击. 该时刻的平行坐标图如图 4 所示, 此时目的端口熵趋近于 1, 说明遭到了多端口扫描攻击, 对应

的源 IP 熵不大, 说明参与扫描的源主机并不多.

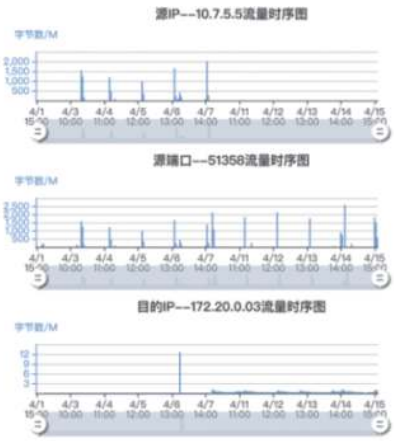

图 7 端口扫描攻击—流量时序图

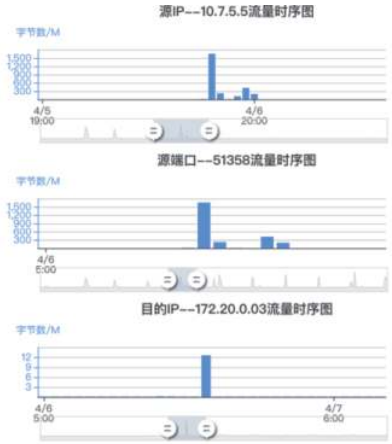

图 8 端口扫描攻击—流量时序图

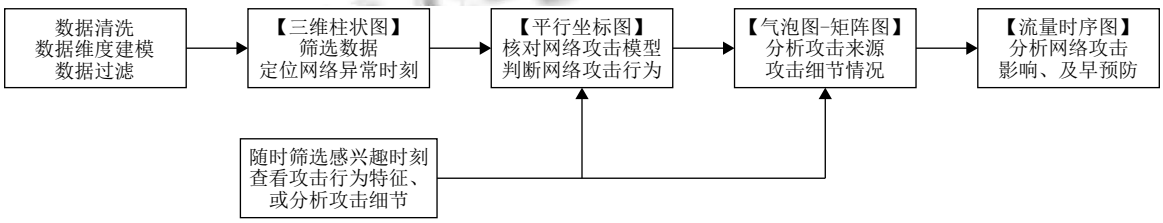

图 9 可视化系统分析流程图

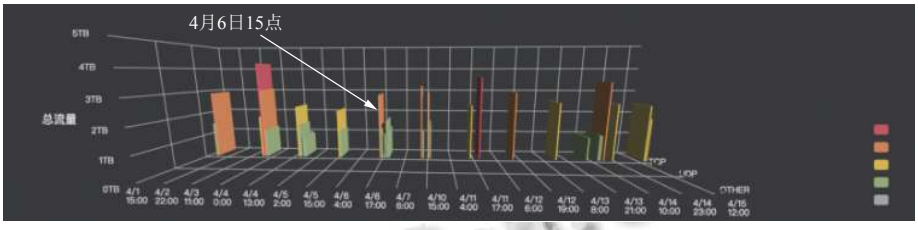

图 10 三维柱状图

为了进一步确定此次攻击中的源主机和目的主机, 我们从矩阵图开始分析, 如图 5 所示, 颜色较深的几个源 IP 地址流量显著异常, 右侧表格中列出了几台源主机的 IP 地址和它们的流量, 再看源端口矩阵图, 除了常用的 80 端口外, 几个非常规端口 (51358, 51357, 45032 和 45031) 的流量异常多.

因为多端口扫描攻击, 必然会使目的 IP 活跃端口数异常增多, 为了确定被攻击的目的 IP, 选择该时刻下的目的 IP 连接数和目的 IP 活跃端口数, 如图 6, 以目的 IP172.20.0.3 为例, 发现该主机的 6 万多个端口被访问了, 同时发现还有十几台主机也遭到了类似的攻击.

为了更精确的观测源 IP 和被攻击的目的 IP 的流量变化, 我们选择攻击者 10.7.5.5, 发起攻击的源端口 51358 和被攻击者 172.20.0.3 作为流量时序图观测对象, 如图 7 所示是 3 个主体的流量时序概览图, 我们发现源端口 51358 的流量分布有周期性特征, 可能经常作为攻击行为的端口, 安全人员需要对此重点监测. 目的 IP172.20.0.3 仅在 4 月 6 日 19 点有一次异常流量峰值, 源 IP10.7.5.5 在第一周内流量有多次异常值, 说明该源 IP 曾发出过不止一次攻击行为. 图 8 是我们根据时间坐标轴定位到 4 月 6 日 19 点时刻左右, 近距离观察该时刻的攻击行为, 从源 IP10.7.5.5 和源端口 51358 的流量时序图, 可以发现此次攻击行为约从 4 月 6 日 15 点开始持续到 20 点左右结束.

综上, 2013 年 4 月 6 日 15 点开始到 4 月 6 日 20 点左右, 内部网络遭受了外部攻击, 在 19 点左右, 10.7.5.5 等多台主机对监控网络的多台主机发起了多端口扫描攻击, 约 6 万多个端口被访问了, 网络安全人员可以根据该可视化图形及时做出响应, 封锁相关的源 IP 和源端口, 降低攻击的影响.

3.2 DDOS 攻击可视化分析

DDOS 攻击是将多个傀儡机联合起来作为攻击平台, 对一个或多个目标发动 DOS 攻击, 成倍地提高拒绝服务攻击的威力, 从而导致目标迅速瘫痪. 如果在很短的时间内, 有大量的不同源 IP 朝同一目的 IP 大量发包, 就表示 DDoS 攻击存在, 匹配表 1 异常熵模式图, 会发现 DDOS 攻击会造成源 IP 熵增多, 目的 IP 熵减少. 依据此特征, 我们选择高亮 4 月 2 日 14 点, 如图 11, 发现该时刻下, 连接数和字节数较多, 但目的 IP 熵和目的端口熵趋近于 0, 说明活跃的目的主机较少, 同时源 IP 熵不小, 源端口熵趋近于 1. 表示此次攻击行为是联合多台主机通过多个端口集中针对某些 IP 发起的, 符合 DDOS 攻击的行为特征.

该时刻下 Top-20 流量矩阵图如图 12, 我们发现颜色较深的方块较多, 表示异常活跃的源 IP 数较多, 右侧列表表示源端口 80、0、123 端口等流量较多, 说明此次攻击行为是多主机借助一般常用端口发起的攻击行为. 活跃的目的 IP 有多个, 其中目的 IP172.30.0.4 流量显著异常, 结合气泡图的目的 IP 连接数发现, 如图 13, 该时刻下目的 IP10.6.6.6, 10.6.6.14, 172.30.0.4 等承受了大量的攻击, 其中目的 IP172.30.0.4 连接数达到 700 多万个, 同时多台源 IP 的连接数都表现异常, 表明这些主机都参与了此次 DDOS 攻击.

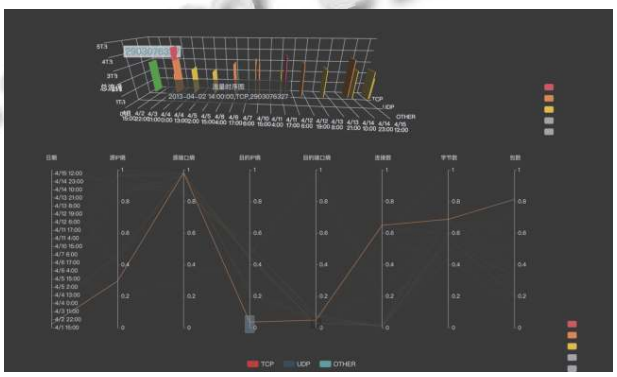

图 11 DDOS 攻击—概览图

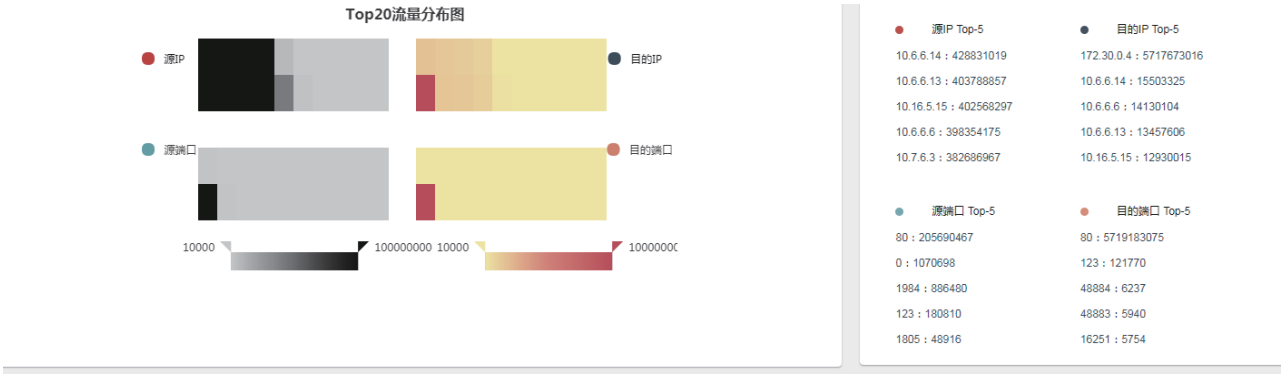

图 12 DDOS 攻击—矩阵图

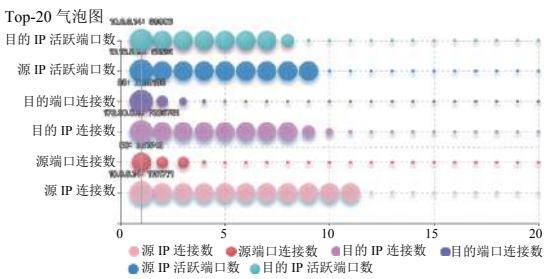

图 13 DDOS 攻击—气泡图

我们选择部分较活跃的攻击方源 IP10.6.6.14、10.16.5.15, 受攻击方目的 IP172.30.0.4 作为此次分析的入口, 观测他们在其他时刻的流量情况. 如图 14 所示, 发现攻击方 10.6.6.14 和 10.16.5.15 在两周内只出现过一次流量高峰, 但受害服务器 172.30.0.4 分别在 4 月 3 日 19 点左右和 4 月 11 日 20 点左右还有另外两次流量高峰期, 可以推测该服务器在这两个时间段内有其他异常情况出现.

综上, 2013 年 4 月 2 日 14 点左右网络受到了 DDOS

攻击, 该次攻击是联合多台源主机针对特定的某几个主机发起的, 如 172.30.0.4、10.6.614 等, 受攻击的端口主要集中在常用的 80、123 端口, 同时目的 IP172.30.0.4 在其他时刻还有疑似的异常行为.

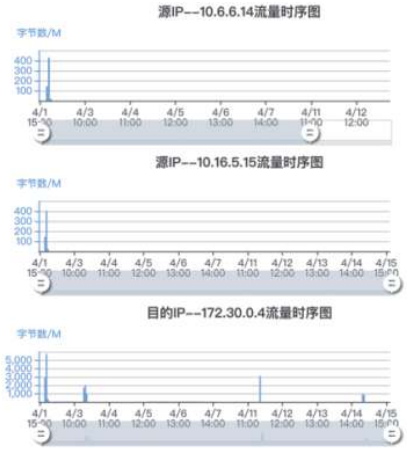

图 14 DDOS 攻击—流量时序图

4 结束语

网络可视化技术是未来网络安全监测的重要发展方向和研究热点. 本文构建的网络可视化系统采用多图联动的形式, 优化了高维度数据的可视化, 为用户提供了丰富的可交互式操作, 有利于用户发掘数据间的关联, 提升用户体验. 但本系统采用的数据类型较为单一, 而且不能实时监测网络安全数据, 无法帮助安全人员及早发现网络威胁. 本文的下一步研究方向, 将引入多种类型的网络安全数据, 采用数学模型提取各种数据的特征, 利用 Spark 对数据的快速处理和实时展现技术, 构建一个完善的网络安全评估系统.

参考文献

1 Nunnally T, Chi P, Abdullah K, *et al*. P3D: A parallel 3D coordinate visualization for advanced network scans. Proceedings of 2013 IEEE International Conference on Communications (ICC). Budapest, Hungary. 2013. 2052–2057.

2 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. [doi: 10.1007/s00607-013-0286-4]

3 吴亚东, 蒋宏宇, 赵思蕊, 等. 网络安全数据 3D 可视化方法. 电子科技大学学报, 2015, 44(4): 594–598, 604. [doi: 10.3969/j.issn.1001-0548.2015.04.020]

4 陈鹏, 司健, 于子桓, 等. 基于信息熵的网络流异常监测和三维可视方法. 计算机工程与应用, 2015, 51(12): 88–93. [doi: 10.3778/j.issn.1002-8331.1408-0111]

5 张胜, 施荣华, 赵颖. 基于多元异构网络安全数据可视化融合分析方法. 计算机应用, 2015, 35(5): 1379–1384, 1416. [doi: 10.3969/j.issn.1001-3695.2015.05.025]

6 赵颖, 王权, 黄叶子, 等. 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.

7 Zhou FF, Huang W, Zhao Y, *et al*. ENTVis: A visual analytic tool for entropy-based network traffic anomaly detection. IEEE Computer Graphics and Applications, 2015, 35(6): 42–50. [doi: 10.1109/MCG.2015.97]

8 Shi L, Wang C, Wen Z, *et al*. 1.5D Egocentric dynamic network visualization. IEEE Transactions on Visualization and Computer Graphics, 2015, 21(5): 624–637. [doi: 10.1109/TVCG.2014.2383380]

9 VAST challenge homepage. http://www.vacommunity.org/VAST+Challenge+2013, 2013.