**2018年第五届中国可视化与可视分析大会**

**数据可视分析挑战赛 挑战1**

**（****ChinaVis Data Challenge 2018 mini challenge 1）**

## 背景介绍

HighTech是一家互联网高科技公司，有几百名员工，分属**财务、人力资源和研发**三个部门。公司正在全力研发一款重量级新产品，近期该产品临近发布，公司对内部发生的一切异常现象都非常敏感。为了维护公司的核心利益，确保新产品顺利发布，公司高层决定临时成立内部威胁情报分析小组，该小组将根据公司内部采集到的数据，分析并处置可能存在的各种安全威胁。在分析威胁情报过程中，数据的复杂性需要计算智能处理，但发现和处置安全威胁需要人的经验、认知和判断，可视分析技术能将计算智能与人类智慧紧密结合，是威胁情报人员高效分析和理解威胁情报数据的利器。假设您是威胁情报分析小组的成员，请您设计并实现一套可视分析解决方案，帮助该公司及时准确地找出可能存在的内部威胁情报。

## 数据说明

本次比赛提供了该公司内部2017年11月共30天的多种监控数据，包括登录日志、网页访问日志、邮件日志、打卡日志和TCP流量日志。数据以csv格式按天分类给出，总共120MB（未压缩之前）。各类型日志详细说明如下：

**登录日志：**员工通过自己主机或跳板机的应用程序，登录服务器或数据库时生成的日志。例如使用SSH、SCP命令、XSHELL程序或者SFTP传输文件都会产生远程登录日志；客户端应用程序访问数据库时，会产生数据库登录日志。

|  |  |  |
| --- | --- | --- |
| **login.csv** | | |
| **字段名称** | **字段含义** | **相关说明** |
| time | 日志生成时间 |  |
| user | 用户名 | 登录使用的用户名 |
| proto | 应用的协议 | 例如ssh、mysql等 |
| dip | 目的IP | 被登录IP |
| dport | 目的端口 | 被登录端口 |
| sip | 源IP | 登录发起IP |
| sport | 源端口 | 登录发起端口 |
| state | 登录结果 | 成功或者失败 |

**网页访问日志：**该日志记录了公司内部所有员工的网页访问记录。time是该条记录生成时间，sip是客户端IP，sport是客户端端口，dip是服务器IP，dport是服务器端口，host是服务器域名。如果通过IP地址直接访问网站，不需要DNS 服务器解析，则HTTP报头的host字段为空字符串。

|  |  |  |
| --- | --- | --- |
| **weblog.csv** | | |
| **字段名称** | **字段含义** | **相关说明** |
| time | 日志生成时间 |  |
| sip | 源IP | 客户端IP |
| sport | 源端口 | 客户端应用端口 |
| dip | 目的IP | 服务端IP |
| dport | 目的端口 | 服务端应用端口 |
| Host | 请求的域名 | HTTP报头的host字段 |

**TCPLOG日志：**记录公司内部网络活动产生的TCP连接。stime、dtime分别是连接建立和断开时间。proto是IP包头中的协议字段值。sip、dip分别是连接发起者和接受者的IP地址，sport、dport是与之对应的源与目的端口。整个连接过程中，sip向dip发送的总字节数为uplink\_length，downlink\_length与之相反。员工的登录行为、网页访问行行为、邮件发送或者接收行为等都会产生一条或者多条TCPLOG日志。

|  |  |  |
| --- | --- | --- |
| **tcpLog.csv** | | |
| **字段名称** | **字段含义** | **相关说明** |
| stime | TCP数据流开始时间 | TCP流的开始时间，即收到该流的第一个SYN包的时间 |
| dtime | TCP数据流结束的时间 | TCP流的结束时间，即收到该流的最后一个包的时间 |
| proto | 协议 | IP包头中的协议字段值 |
| dip | 目的IP | TCP 数据流的服务端IP |
| dport | 目的端口 | TCP 数据流的服务端应用端口 |
| sip | 源IP | TCP 数据流的客户端发起IP |
| sport | 源端口 | TCP 数据流的客户端应用端口 |
| uplink\_length | 上行字节数 | 从TCP流的建立到该流的结束，从客户端发往服务器端的应用层数据的字节总数 |
| downlink\_length | 下行字节数 | 从TCP流的建立到该流的结束，从服务器端发往客户端的应用层数据的字节总数 |

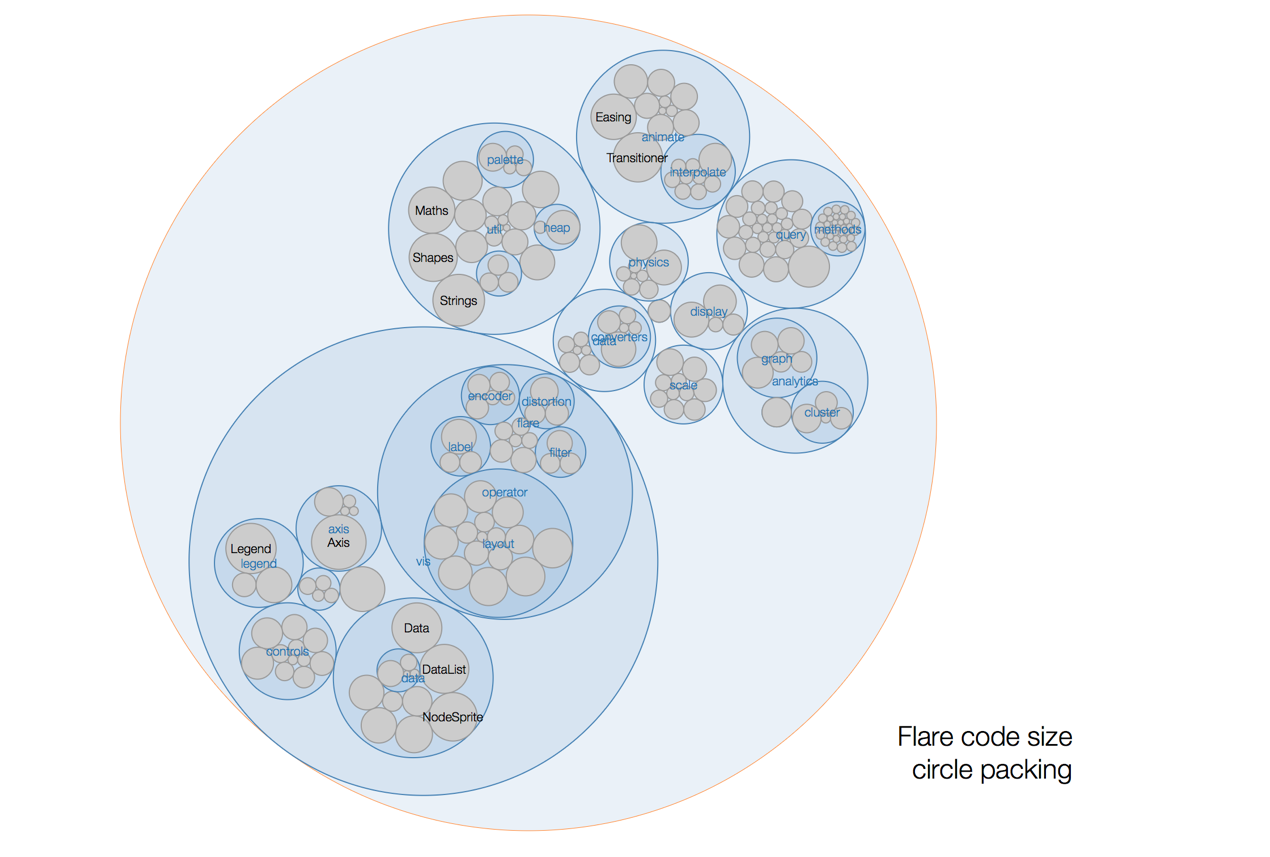
**邮件日志：**邮件日志记录了经过公司邮件服务器的收发邮件信息。time是邮件的发送时间/接收时间，proto是邮件使用的应用协议。sip、dip分别是连接发起者和接受者的IP地址，sport、dport是与之对应的源与目的端口。from、to分别是邮件的发送者和接收者。邮件内容属于隐私，只提供邮件主题subject。

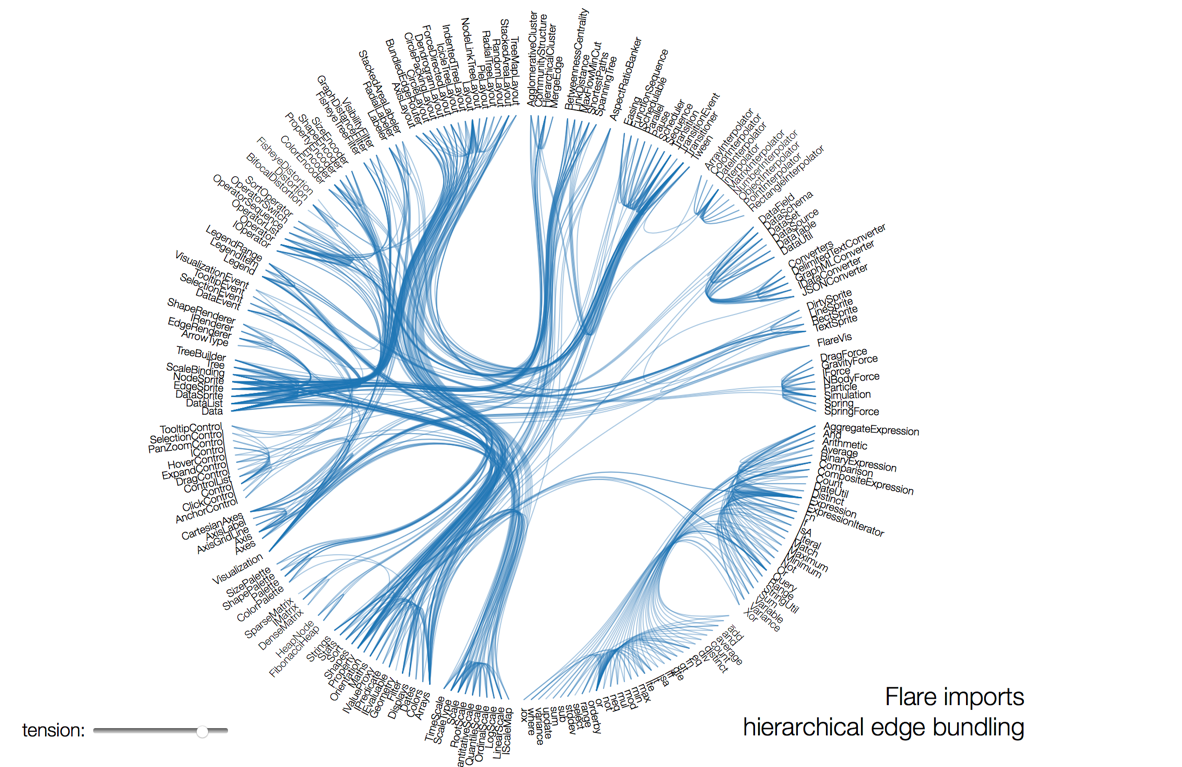
|  |  |  |
| --- | --- | --- |
| **email.csv** | | |
| **字段名称** | **字段含义** | **相关说明** |
| time | 邮件发送/接收时间 | 邮件包头中的邮件发送/接收时间 |
| proto | 应用协议 | SMTP |
| sip | 源IP | IP报头源IP地址 |
| sport | 源端口 | TCP报头源应用端口 |
| dip | 目的IP | IP报头目的IP地址 |
| dport | 目的端口 | TCP报头目的应用端口 |
| **from** | **邮件发送人** | **来自于邮件头相应字段** |
| **to** | **邮件接收人** | **来自于邮件头相应字段，出现多个接收人时用分号隔开。** |
| **subject** | **主题** | **来自于邮件头相应字段** |

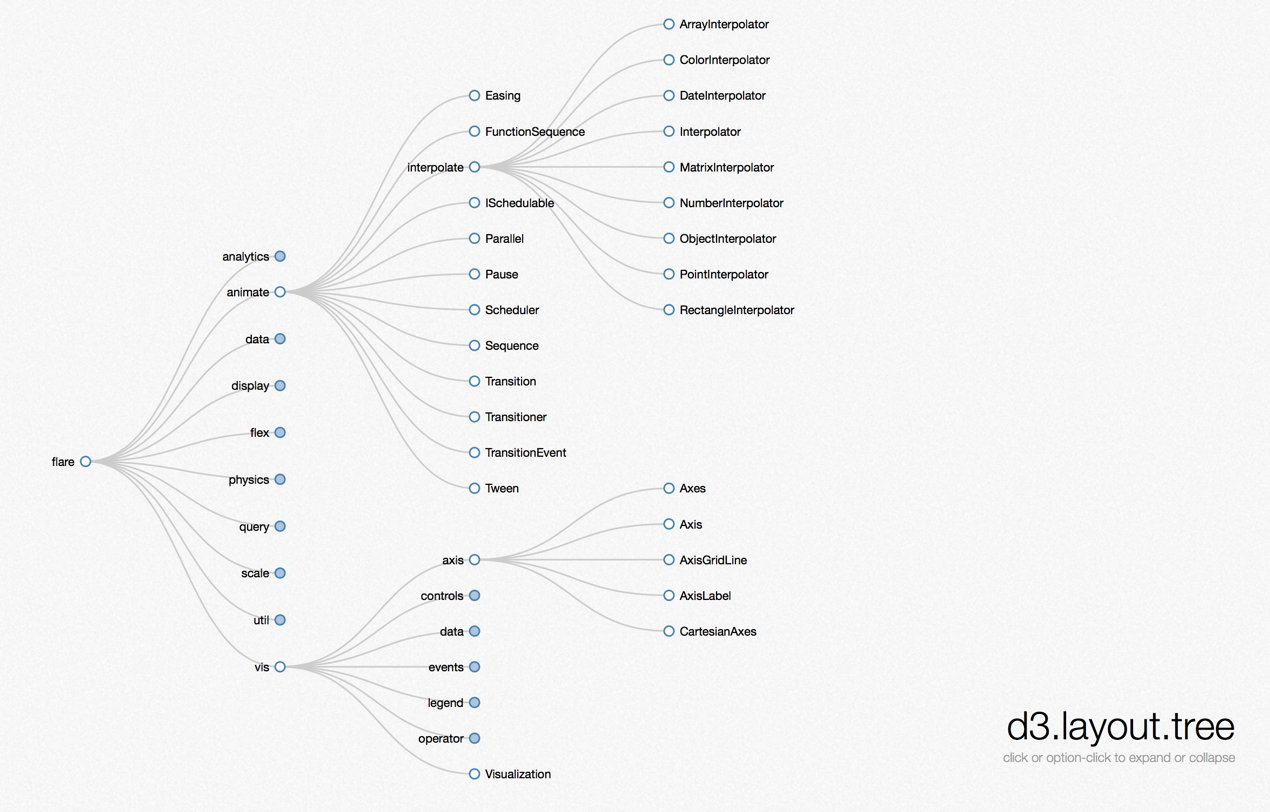
**打卡日志：**记录了公司每个员工每天上下班时间，一行记录中 checkin和checkout都为0，表示没来上班。那就是说，没有来公司打卡的员工，也会生成一行考勤记录。另外，如果公司员工当天没来公司上班，次日该员工会收到旷工提醒邮件。

|  |  |  |
| --- | --- | --- |
| **checking.csv** | | |
| **字段名称** | **字段含义** | **相关说明** |
| id | 员工id |  |
| day | 日期 |  |
| checkin | 上班签到时间 |  |
| checkout | 下班签退时间 |  |

## 题目说明

1. 分析公司内部员工所属部门及各部门的人员组织结构，给出公司员工的组织结构图（建议参赛者回答此题文字不多于500字，图片不多于5张）；





（2）分析该公司员工的日常工作行为，按部门总结和展示员工的正常工作模式（建议参赛者回答此题文字不多于1000字，图片不多于8张）；

【时序-事件分析】

【时（时间）空（登录、网页浏览记录等）限制 挖掘频繁项集 使用时间事件图的表示方法展现】

（3）找出至少5个异常事件，并分析这些事件之间可能存在的关联，总结你认为有价值的威胁情报，并简要说明你是如何利用可视分析方法找到这些威胁情报的（建议参赛者回答此题文字不多于1500字，图片不多于10张）。

【异常分析：提特征（频率等）找离群点】

考勤、上班时常、连接时间…

（1）统计各字段出现频率