

数据可视分析挑战赛 挑战1

2018年第五届中国可视化与可视分析大会

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。

背景介绍

- 时间：2018/11/01~2017/11/30
- 地点：HighTech公司
- 人物：公司几百名员工（财务、人力资源和研发）
- 起因：公司正在全力研发一款重量级新产品
- 经过：为了维护公司的核心利益，确保新产品顺利发布，公司内部采集数据，用于分析并处置可能存在的各种安全威胁
- 结果：设计并实现一套可视分析解决方案，及时准确地找出可能存在的内部威胁情报

数据说明

公司内部2017年11月共30天的多种监控数据：

打卡日志	公司每个员工每天上下班时间 【所有员工】
邮件日志	经过公司邮件服务器的收发邮件信息 【所有员工】
网页访问日志	所有员工的网页访问记录 【所有员工】
登录日志	员工登录服务器或数据库时生成的日志 【研发部门员工】
TCP流量日志	公司内部网络活动产生的TCP连接 【研发部门员工】

Question

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）
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）

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。
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。

字段名称	字段含义	相关说明
id	员工工号	
day	日期	
checkin	上班签到时间	
checkout	下班签退时间	

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

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 1000(工号)@hightech.com	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

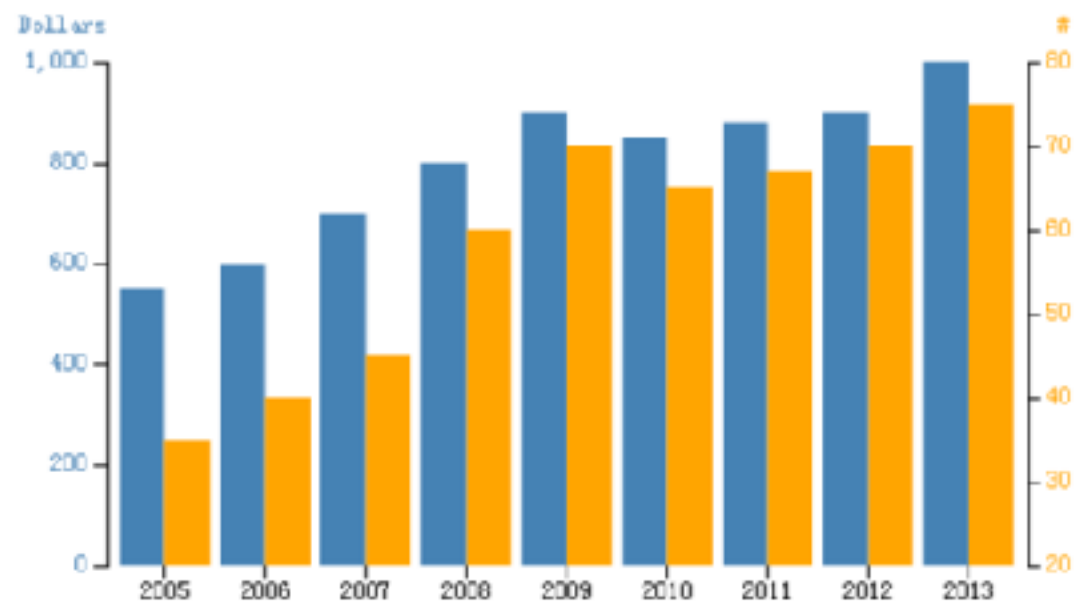
From & To

- 1. 发件人与收件人均为 xxx@hightech.com
 - 邮件主题内容 ==> 部门
 - 监控、安全、需求、网络... ==> 研发部门
 - 税务、成本、财务、报销... ==> 财务部门
 - 录用、福利、照片... ==> 人力资源部门
 - 收发件人 ==> 部门组织结构
 - 邮件群发 ==> 部门组织结构
 - 监控报警
 - 公司福利
 - ...

	研发部门	财务部门	Hr部门
研发部门	工作报告 监控报警 ...	报销 ...	面试 招聘 ...
财务部门	报销 ...	工作汇报 财务报表 ...	工作材料 报销 ...
Hr部门	团建福利 面试招聘 ...	团建福利 面试招聘 ...	工作汇报 工作材料 ...

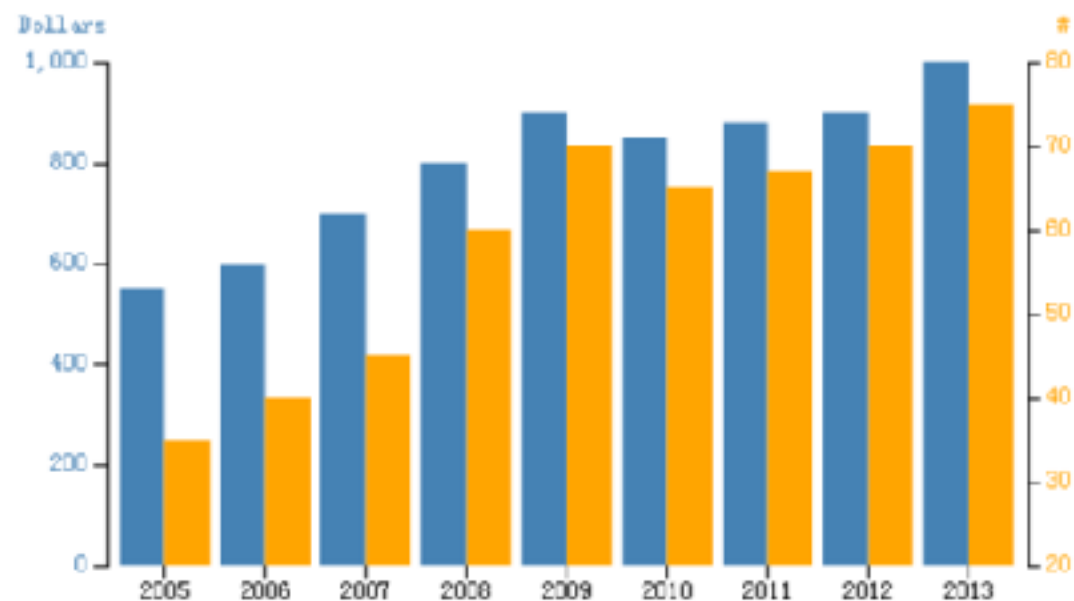
From & To

- 2. 发件人为xxx@hightech.com, 收件人为外部邮箱 @126 @qq...
 - 邮件主题内容
 - 录用 ==> 人力资源部门
 - 公司介绍 ==> 人力资源部门
 - 合作
 - 泄密威胁
 - ...



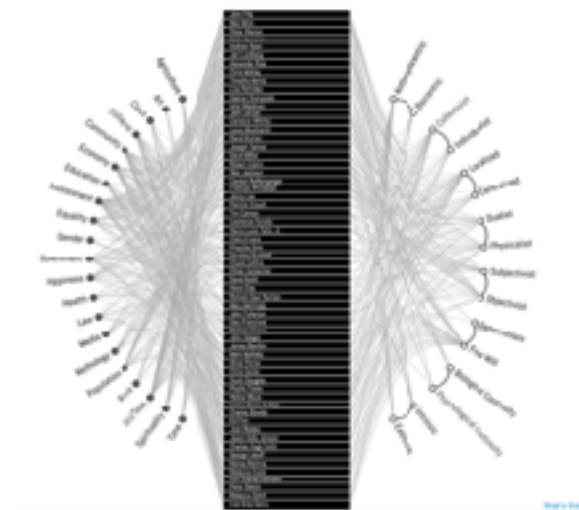
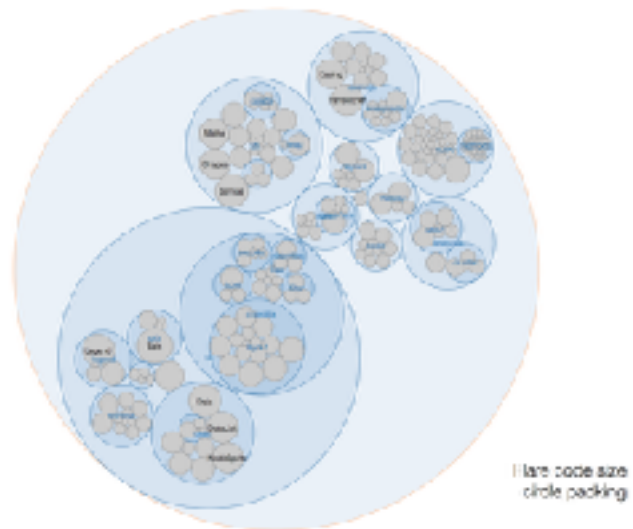
From & To

- 3. 收件人为xxx@hightech.com, 发件人为外部邮箱
 - 邮件主题内容
 - 广告
 - 合作
 - 猎头推荐职位
 - 威胁
 - ...



From 邮件日志

- 1. 员工工号与员工所使用IP的映射
 - 2. 员工所属部门
 - 3. 部门结构
 - 4. 正常工作模式—邮件
-
- 难点：
 - 如何得到部门的层级组织网络
 - 如何更好的展示部门结构图



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

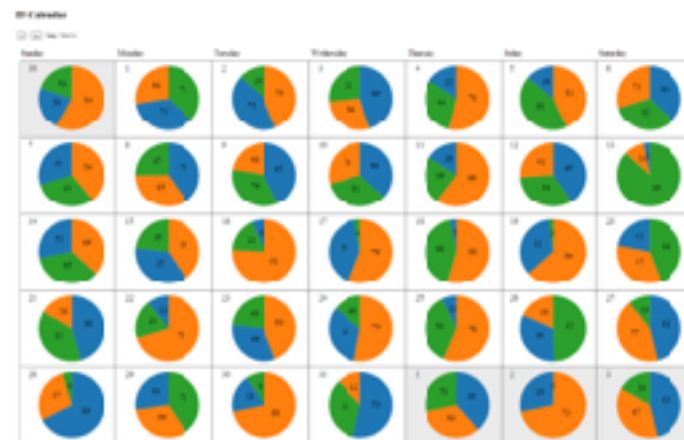
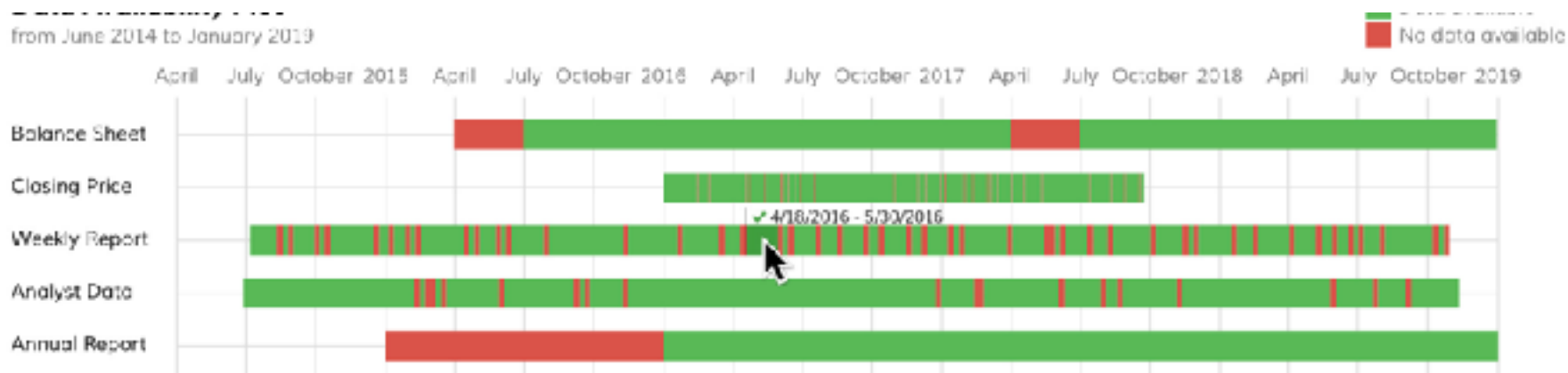
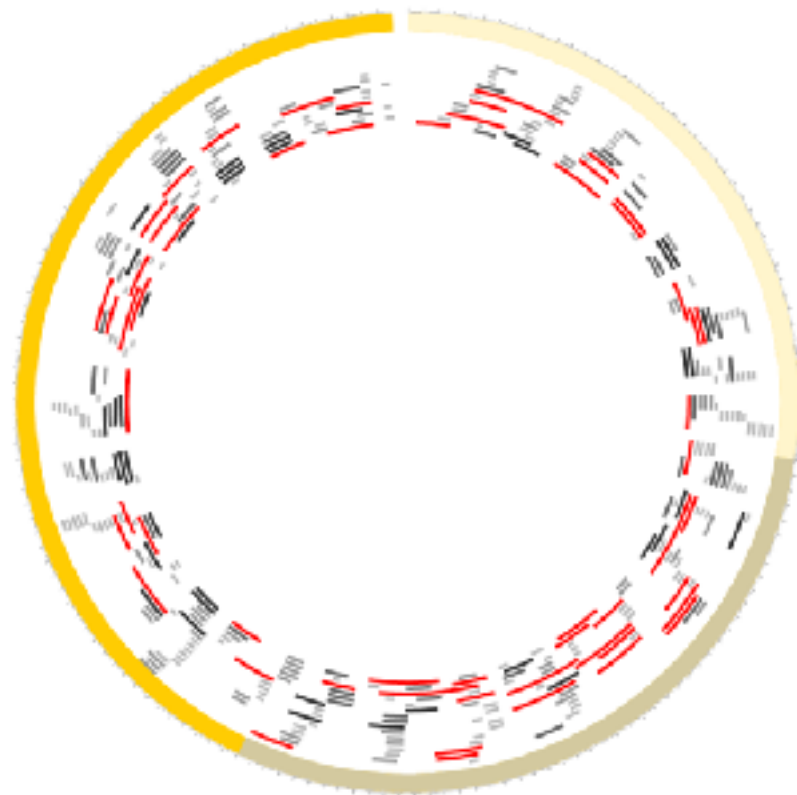
字段名称	字段含义	相关说明
time	日志生成时间	
sip	源IP 【确定员工个人】	客户端IP
sport	源端口	客户端应用端口
dip	目的IP	服务端IP
dport	目的端口	服务端应用端口
Host	请求的域名	HTTP报头的host字段

From 网页访问日志

- 得到员工工作模式
 - 对域名进行分类 => 正常工作模式
 - 工作：域名中带有hightech，例如oa、git、email、lib...
 - 工作：搜索引擎，baidu、google...
 - 摸鱼：Taobao、novel、store、tudou、cntv、games...
 - 刚上班
 - 午休
 - 快下班时
 - 域名频率 => 访问频率过低 => 异常
 - 刷新频率 => 刷新频率过快 => 异常
 - ...

From 网页访问日志

- 难点:
 - 1. 域名分类如何更加精准
 - 2. 展示时间-事件???



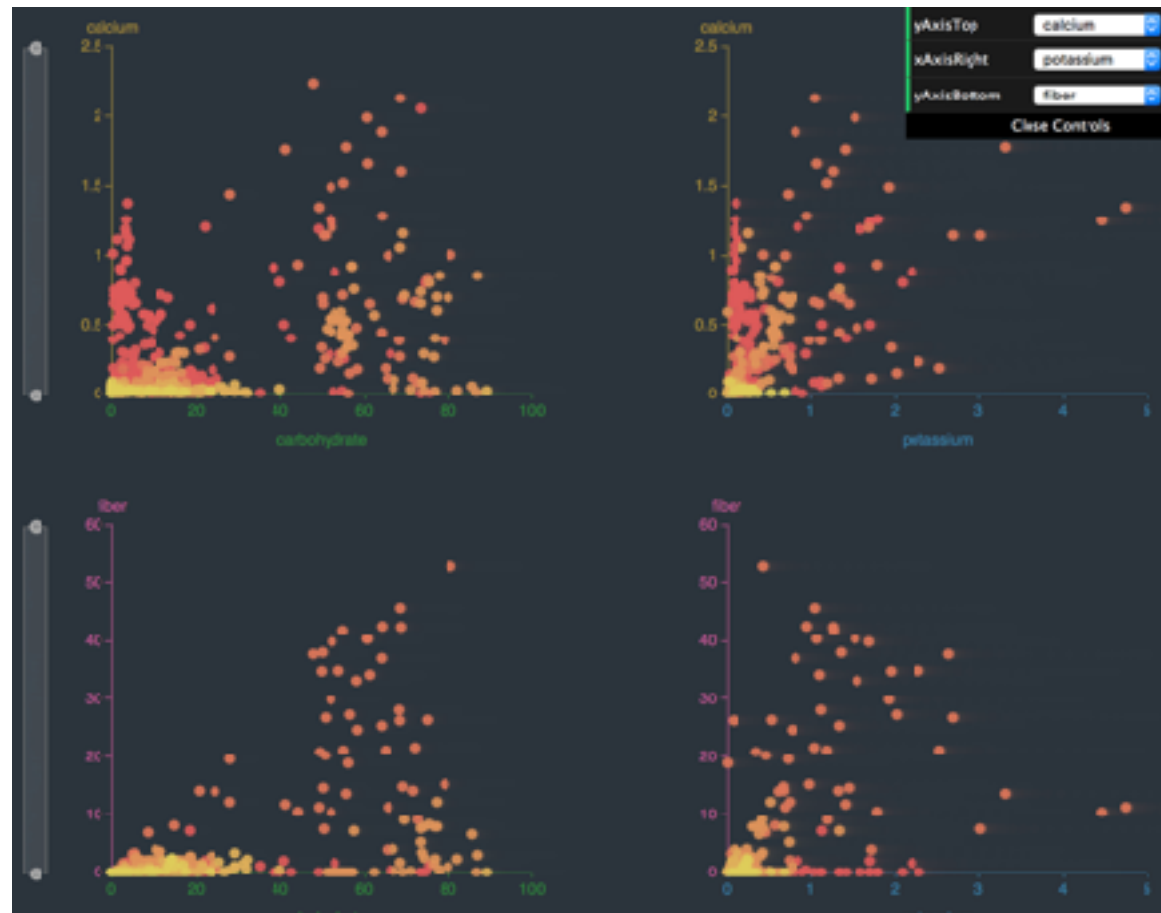
登录日志

员工登录服务器或数据库时生成的日志

字段名称	字段含义	相关说明
time	日志生成时间	
user	用户名【root、各组使用的用户名】	登录使用的用户名
proto	应用的协议【代表使用目的】	例如ssh、mysql、scp等
dip	目的IP【服务器IP地址】	被登录IP
dport	目的端口	被登录端口
sip	源IP【确定员工个人】	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

From 登录日志

- 1. ip映射为工号 + user登录名
 - => 程序员分组
- 2. proto + time
 - => 程序员工作模式
- 3. Proto + time + states
 - => 威胁 or 正常



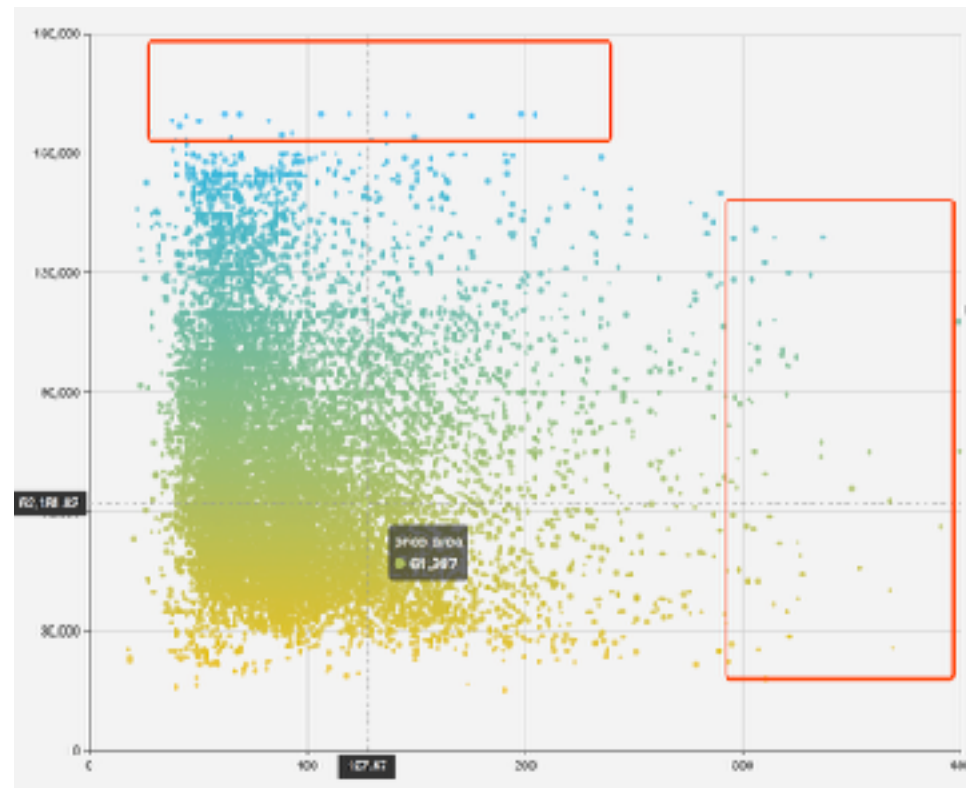
TCPLOG日志

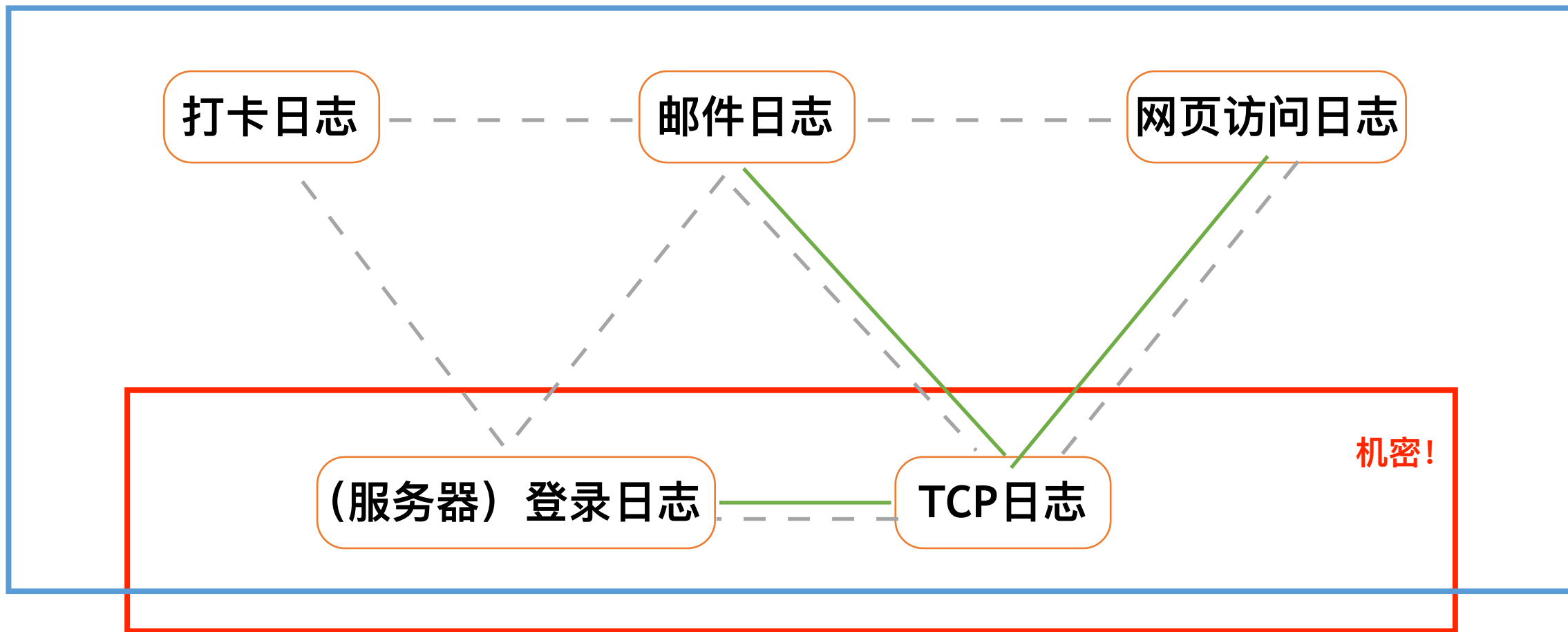
记录公司内部网络活动产生的TCP连接，员工的登录行为、网页访问行行为、邮件发送或者接收行为等都会产生一条或者多条TCPLOG日志。

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
dtype	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，ssh, ftp, smtp, mysql...
dip	目的IP	TCP 数据流的服务端IP 【服务器IP地址、员工PC的IP地址、外部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 【服务器IP地址】
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

From TCPLLOG 日志

- 1. 对proto分类, smtp、ssh、mysql、ftp...
 - **注意! 能够泄露消息的proto!**
- 2. 对每个类别抽取多个特征, 两两绘制散点图, 找出异常点
 - Tcp连接时长 - Tcp开始时间
 - Tcp连接时长 - 上行字节数
 - Tcp连接时长 - 下行字节数
 - Tcp开始时间 - 上行字节数
 - Tcp开始时间 - 下行字节数
 - ...
 - **多试些特征、总会有离群点出现的!**





讨论

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）
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）

进展汇报 20180420

回顾

- Question 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）

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

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 1000(工号)@hightech.com	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

From & To

- 1. 发件人与收件人均为 xxx@hightech.com
 - 邮件主题内容 ==> 部门
 - 监控、安全、需求、网络... ==> 研发部门
 - 税务、成本、财务、报销... ==> 财务部门
 - 录用、福利、照片... ==> 人力资源部门
 - 收发件人 ==> 部门组织结构
 - 邮件群发 ==> 部门组织结构
 - 监控报警
 - 公司福利
 - ...

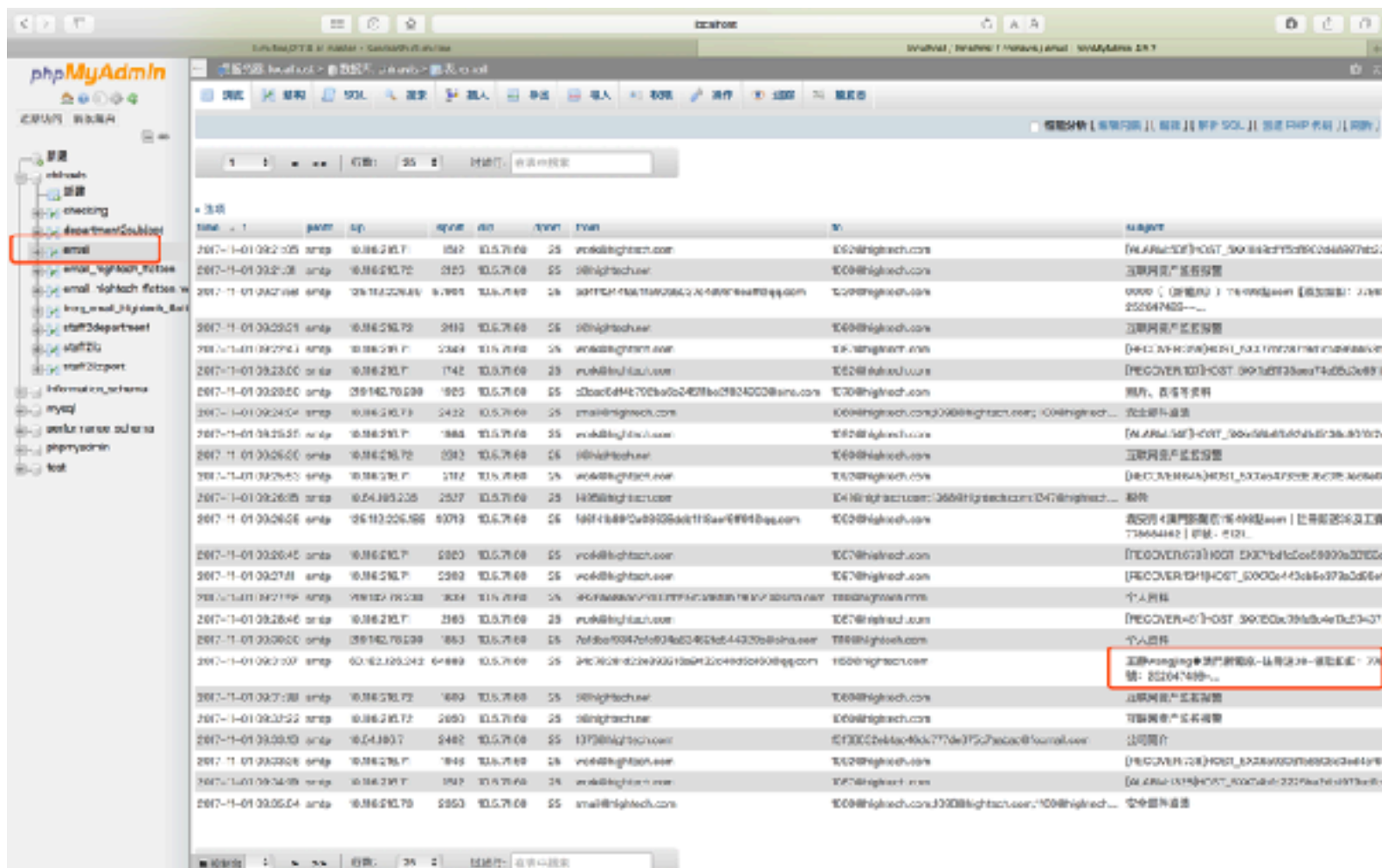
	研发部门	财务部门	Hr部门
研发部门	工作报告 监控报警 ...	报销 ...	面试 招聘 ...
财务部门	报销 ...	工作汇报 财务报表 ...	工作材料 报销 ...
Hr部门	团建福利 面试招聘 ...	团建福利 面试招聘 ...	工作汇报 工作材料 ...

实施

1. 处理数据：将30天数据存入数据库

1. 字符转码

2. 无法解析的字符



id	time	ip	email	subject
1	2017-11-01 09:27:00	10.10.10.10	wangwang@163.com	互联网广告营销
2	2017-11-01 09:27:01	10.10.10.10	wangwang@163.com	互联网广告营销
3	2017-11-01 09:27:02	10.10.10.10	wangwang@163.com	互联网广告营销
4	2017-11-01 09:27:03	10.10.10.10	wangwang@163.com	互联网广告营销
5	2017-11-01 09:27:04	10.10.10.10	wangwang@163.com	互联网广告营销
6	2017-11-01 09:27:05	10.10.10.10	wangwang@163.com	互联网广告营销
7	2017-11-01 09:27:06	10.10.10.10	wangwang@163.com	互联网广告营销
8	2017-11-01 09:27:07	10.10.10.10	wangwang@163.com	互联网广告营销
9	2017-11-01 09:27:08	10.10.10.10	wangwang@163.com	互联网广告营销
10	2017-11-01 09:27:09	10.10.10.10	wangwang@163.com	互联网广告营销
11	2017-11-01 09:27:10	10.10.10.10	wangwang@163.com	互联网广告营销
12	2017-11-01 09:27:11	10.10.10.10	wangwang@163.com	互联网广告营销
13	2017-11-01 09:27:12	10.10.10.10	wangwang@163.com	互联网广告营销
14	2017-11-01 09:27:13	10.10.10.10	wangwang@163.com	互联网广告营销
15	2017-11-01 09:27:14	10.10.10.10	wangwang@163.com	互联网广告营销
16	2017-11-01 09:27:15	10.10.10.10	wangwang@163.com	互联网广告营销
17	2017-11-01 09:27:16	10.10.10.10	wangwang@163.com	互联网广告营销
18	2017-11-01 09:27:17	10.10.10.10	wangwang@163.com	互联网广告营销
19	2017-11-01 09:27:18	10.10.10.10	wangwang@163.com	互联网广告营销
20	2017-11-01 09:27:19	10.10.10.10	wangwang@163.com	互联网广告营销
21	2017-11-01 09:27:20	10.10.10.10	wangwang@163.com	互联网广告营销
22	2017-11-01 09:27:21	10.10.10.10	wangwang@163.com	互联网广告营销
23	2017-11-01 09:27:22	10.10.10.10	wangwang@163.com	互联网广告营销
24	2017-11-01 09:27:23	10.10.10.10	wangwang@163.com	互联网广告营销
25	2017-11-01 09:27:24	10.10.10.10	wangwang@163.com	互联网广告营销
26	2017-11-01 09:27:25	10.10.10.10	wangwang@163.com	互联网广告营销
27	2017-11-01 09:27:26	10.10.10.10	wangwang@163.com	互联网广告营销
28	2017-11-01 09:27:27	10.10.10.10	wangwang@163.com	互联网广告营销
29	2017-11-01 09:27:28	10.10.10.10	wangwang@163.com	互联网广告营销
30	2017-11-01 09:27:29	10.10.10.10	wangwang@163.com	互联网广告营销

实施

2. 基本的数据统计

- 1) 共299个员工
- 2) 14个公共账户

特殊邮箱名	邮件主题示例	备注
work	[ALARM:508]HOST_5XX1149cff f5b8902d48927a...	ALERT:XXX RECOVER:XXX 研发组
ti	互联网资产监控报警	ti@hightech.net 只发给员工1060
smail	安全邮件崩溃	1060@hightech.com ; 1098@hightech.com ; 1100@hightech.com ; 1154@hightech.com ; 1191@hightech.com ; 1207@hightech.com ; 1209@hightech.com
school	培训邀请	To allstaff@hightech.com
notice	通知	To allstaff@hightech.com 个人
guanhuai	福利来啦	To allstaff@hightech.com
fuli	生日快乐!!! 送京东卡福利	
hr	内推-招人啦	To allstaff@hightech.com
kaoqin		
alert		1284 1287 数据库 EmergencyDataBaseFatalError!
allstaff		
meeting	全员大会	To allstaff@hightech.com
it	资源申请	
finance	报销	个人

实施

3. 得到员工-ip映射表，为后期tcp等数据做准备

- 1) 一个工号有且只有一个ip地址与其对应
- 2) 一个员工对应多个端口

name	1	sip	sport
1007		10.64.105.171	2383
1007		10.64.105.171	2375
1007		10.64.105.171	2835
1007		10.64.105.171	2016
1013		10.64.106.4	2035
1013		10.64.106.4	2037
1013		10.64.106.4	2069
1013		10.64.106.4	1737
1013		10.64.106.4	1752
1013		10.64.106.4	1754
1013		10.64.106.4	2608
1013		10.64.106.4	2589
1013		10.64.106.4	2344
1013		10.64.106.4	3388
1013		10.64.106.4	3362
1013		10.64.106.4	3395
1013		10.64.106.4	3369
1013		10.64.106.4	1556
1013		10.64.106.4	1541
1013		10.64.106.4	1698
1013		10.64.106.4	2391
1013		10.64.106.4	2406
1013		10.64.106.4	2426
1013		10.64.106.4	2400
1041		10.64.105.89	2426

name	1	sip
work		10.116.216.71
tl		10.116.216.72
small		10.116.216.73
school		10.1.4.17
notice		10.1.4.17
meeting		10.1.4.17
kaoqin		10.1.4.17
it		10.1.4.17
hr		10.1.4.17
guanhuai		10.1.4.17
full		10.1.4.17
finance		10.1.4.17
allstaff		10.1.4.17
alert		10.63.120.70
1500		10.64.106.66
1499		10.64.106.10
1498		10.64.105.235
1497		10.64.105.106
1496		10.64.105.145
1495		10.64.105.210
1494		10.64.105.193
1493		10.64.105.13
1491		10.64.105.213
1490		10.64.105.176
1489		10.64.105.132

实施

4. 过滤出员工发给员工的邮件，找出其所有主题（共61个主题），然后根据关键词，加以先验知识，给出邮件主题类别“研发/人力资源/财务类别/综合”标签

* 保留work、smail、it、alert邮箱，标签为“研发-监控”

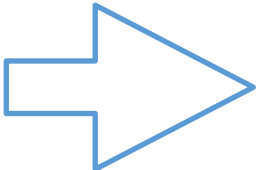
* 不能确定部门的邮件主题标注为“综合”

税务	财务
财务报账	财务
成本控制	财务
资金	财务
财务分析	财务
会计核算	财务
岗位说明书	人力资源
招聘信息总结	人力资源
福利保障	人力资源
劳动合同	人力资源
人员档案	人力资源
考勤相关	人力资源
安全邮件崩溃	研发
地图配置	研发
技术分享安排	研发
总结	研发
传输设置	研发
需求	研发
分析平台配置	研发
项目汇报	研发
终验文档	研发
测试脚本	研发
软件开发文档	研发
需求调研	研发
需求与原型	研发
实施方案	研发
搜索详细设计	研发
项目进度计划	研发
项目过程文档	研发
用户手册	研发

项目计划、问题、风险	研发
api汇总	研发
项目测试数据统计	研发
概要设计	研发
部署文档	研发
前后端接口	研发
初验文档	研发
系统配置子系统	研发
项目周报	研发
项目总结	研发
特殊字段说明	研发
资源申请	研发
月报总结	研发
EmergencyDataBaseFatalError!	研发
【通知】设备已归还，请通过。	研发
工作汇报	综合
例会会议纪要	综合
年度工作目标	综合
近期工作总结	综合
例会	综合
合同	综合
绩效考核	综合
Re:报名参加	综合
工作计划	综合
公司发展规划	综合
年度计划	综合
请假条	综合
事假	综合
请假四天，望批准	综合
【辞职信】	综合
Reply: 辞职请求审核：批准。	综合

实施

5. 处理员工内部邮件数据

1) 原邮件数据（包括群发的邮件）预处理：扁平化处理成一对一形式， 并为其主题添加部门类别标签

2) 以发件人、收件人、邮件类别标签为key统计频率



1500	1096	研发	18
1500	1239	研发	17
1500	1254	研发	17
1500	1368	财务	1
1500	1402	研发	17
1500	1478	研发	17

from	to	subject	topic	1
1498	1041	税务	财务	
1498	1368	税务	财务	
1498	1347	税务	财务	
1498	1255	税务	财务	
1498	1248	税务	财务	
1498	1327	税务	财务	
1498	1439	税务	财务	
1498	1137	税务	财务	
1498	1370	税务	财务	
1498	1467	税务	财务	
1498	1226	税务	财务	
1498	1369	税务	财务	
1498	1188	税务	财务	
1498	1213	税务	财务	
1498	1451	税务	财务	
1498	1124	税务	财务	
1498	1431	税务	财务	
1498	1293	税务	财务	
1498	1253	税务	财务	
1498	1342	税务	财务	
1498	1108	税务	财务	
1498	1180	税务	财务	
1498	1346	税务	财务	
1186	1291	财务报账	财务	
1431	1041	成本控制	财务	

实施

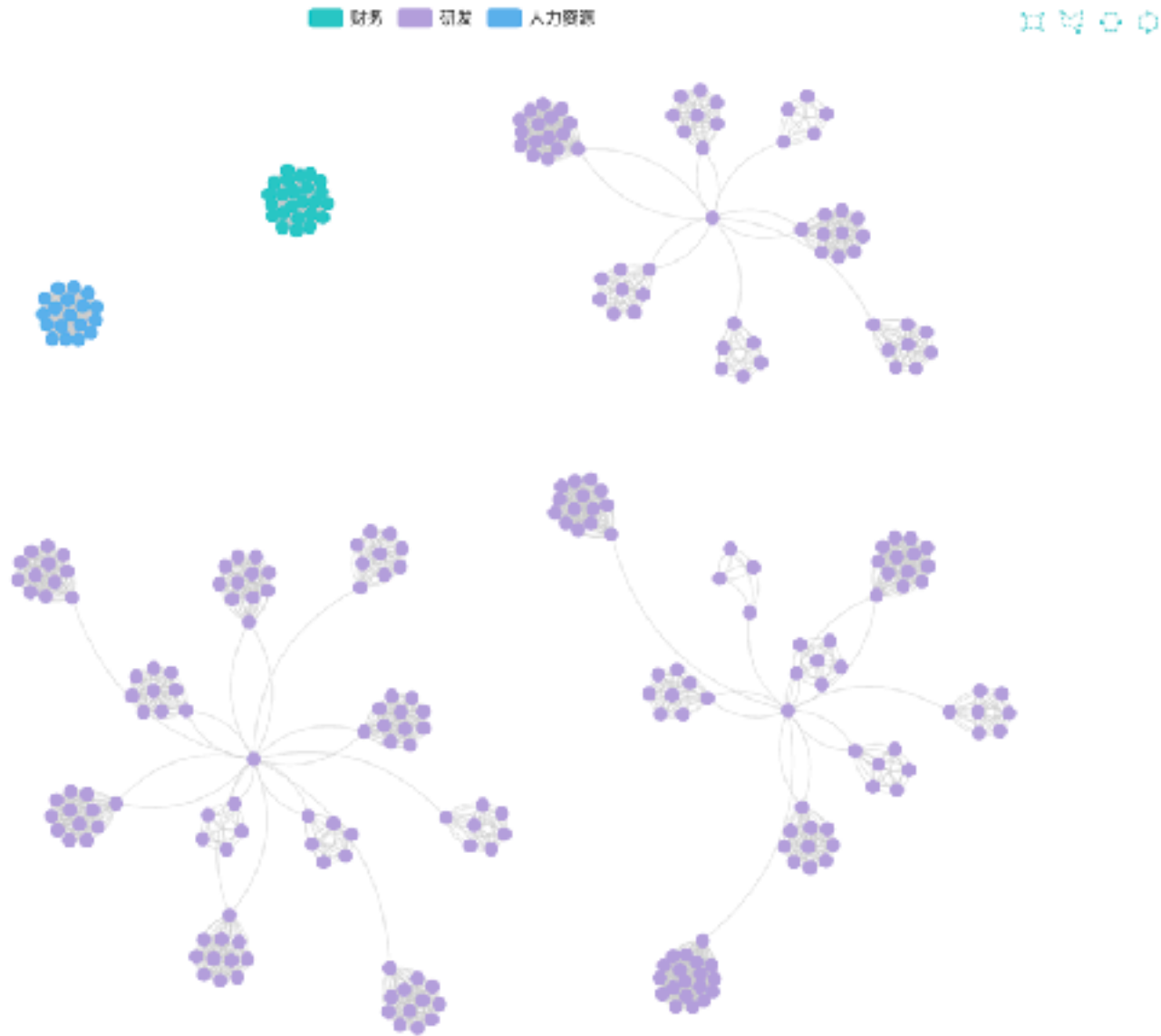
5. 使用[from, to, topic, freq]绘制力导向图

- * 过滤发信频率低于3的数据
- * 过滤“综合”，“研发-监控” 标签的邮件

=====➡

部门结构清晰

部门领导明确



实施

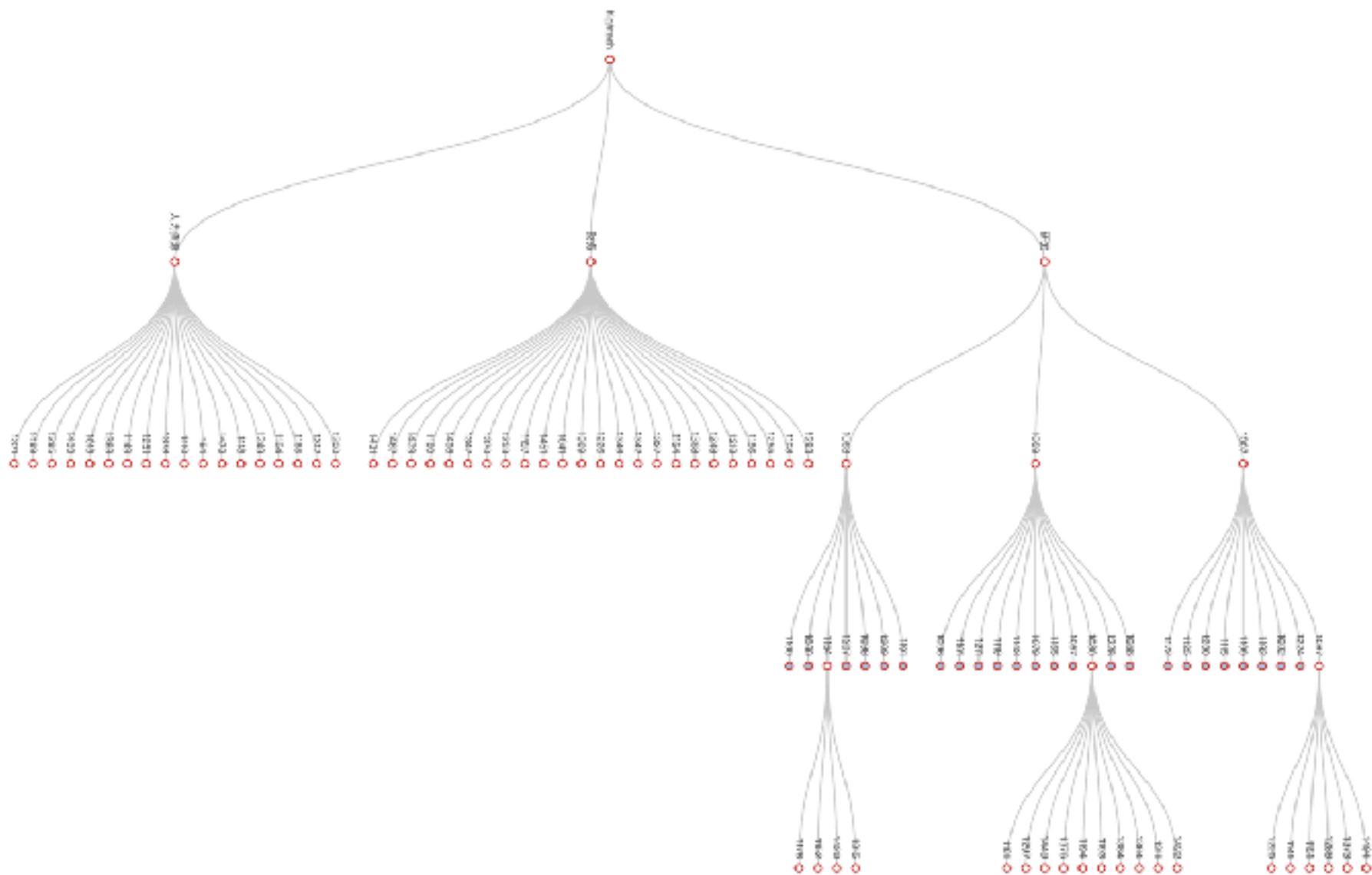
6. 根据力导向图为每个员工划分了所属的部门, 其中财务部门和行政部门结构比较简单, 研发部门又分为三个组, 每个组下又分有10个左右的小组, 将数据存储至数据库

staff	department	staff	department	staff	department
1253	财务	1458	研发_A_1	1296	研发_A_4
1137	财务	1305	研发_A_1	1330	研发_A_4
1451	财务	1474	研发_A_1	1189	研发_A_4
1041	财务	1405	研发_A_1	1103	研发_A_4
1369	财务	1170	研发_A_1	1263	研发_A_4
1226	财务	1362	研发_A_1	1319	研发_A_4
1346	财务	1481	研发_A_1	1399	研发_A_4
1342	财务	1493	研发_A_1	1098	研发_A_5_loader
1327	财务	1379	研发_A_1	1127	研发_A_5
1124	财务	1060	研发_A_2_leader	1277	研发_A_5
1368	财务	1328	研发_A_2	1334	研发_A_5
1248	财务	1446	研发_A_2	1343	研发_A_5
1213	财务	1306	研发_A_2	1496	研发_A_5
1186	财务	1457	研发_A_2	1209	研发_A_6_leader
1255	财务	1440	研发_A_2	1153	研发_A_6
1108	财务	1145	研发_A_2	1460	研发_A_6
1293	财务	1396	研发_A_2	1126	研发_A_6
1068	研发_A_leader	1359	研发_A_2	1339	研发_A_6
1100	研发_A_1_leader	1336	研发_A_2	1349	研发_A_6
1321	研发_A_1	1154	研发_A_3_leader	1388	研发_A_6
1159	研发_A_1	1176	研发_A_3	1322	研发_A_6
1385	研发_A_1	1152	研发_A_3	1191	研发_A_7_leader
1139	研发_A_1	1420	研发_A_3	1428	研发_A_7
1234	研发_A_1	1315	研发_A_3	1469	研发_A_7
1147	研发_A_1	1207	研发_A_4_leader	1156	研发_A_7

实施

7. 绘制

树形图



实施

8. 划分小组职能

通过聚类图得到小组的划分，那么再去以小组为单位，统计每个组的邮件主题频数，又因为每个组的规模不同，频数无法估量每个主题的重要程度，统计每组每个主题的频率

subject	freq	department
分析平台配置	285	研发_A_1
需求与原型	240	研发_A_1
api汇总	240	研发_A_1
软件开发文档	210	研发_A_1
特殊字段说明	210	研发_A_1
需求调研	195	研发_A_1
初验文档	195	研发_A_1
概要设计	195	研发_A_1
测试脚本	195	研发_A_1
项目过程文档	180	研发_A_1
传输设置	180	研发_A_1
地图配置	150	研发_A_1
需求	150	研发_A_1
项目测试数据统计	150	研发_A_1
部署文档	135	研发_A_1
系统配置子系统	135	研发_A_1
前后端接口	135	研发_A_1
终验文档	120	研发_A_1
实施方案	105	研发_A_1
搜索详细设计	105	研发_A_1
用户手册	90	研发_A_1
项目计划、问题、风险	60	研发_A_1
例会会议纪要	30	研发_A_1
总结	27	研发_A_1
近期工作总结	22	研发_A_1

实施

9. 划分小组职能

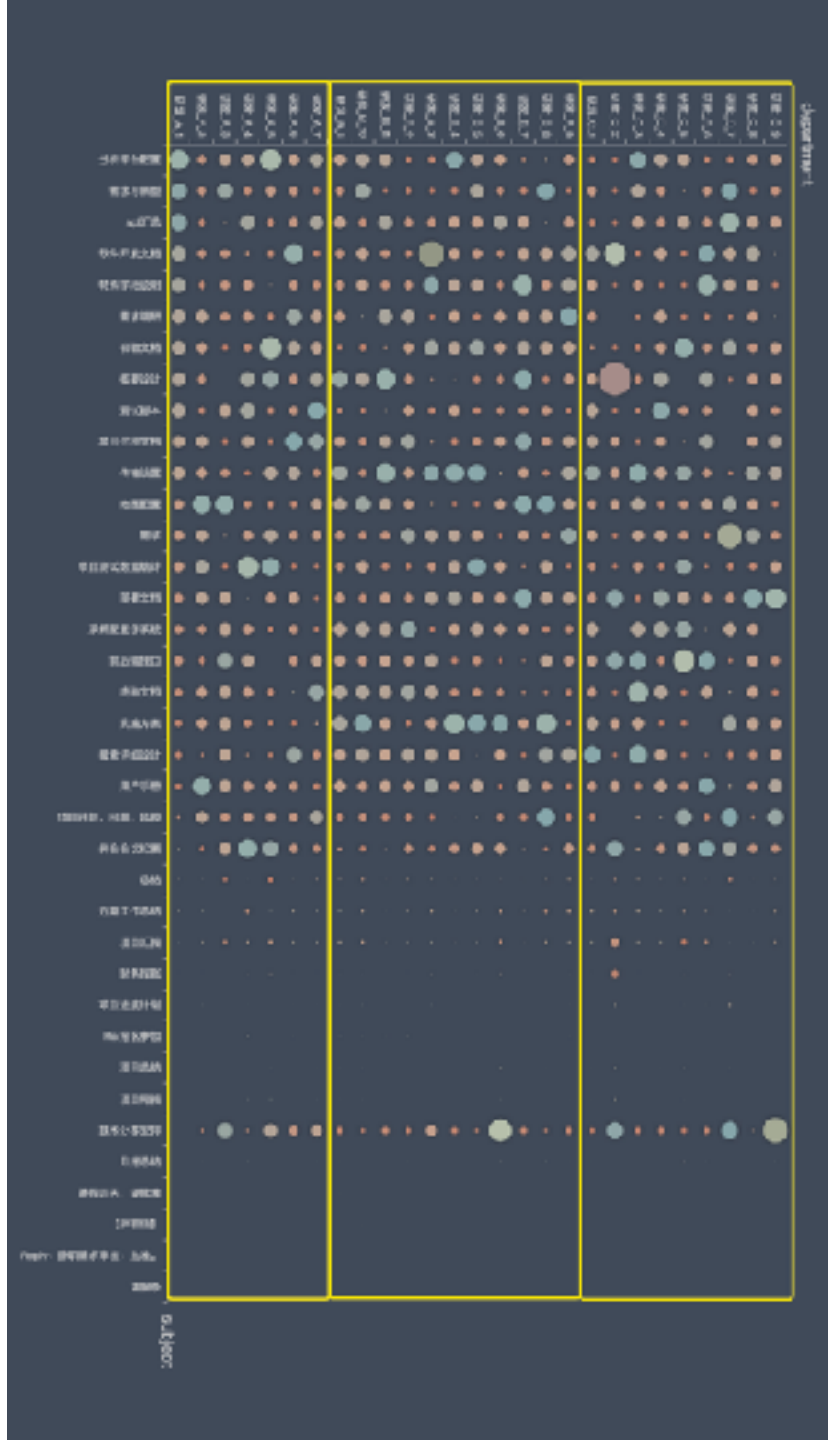
根据每组每个邮件主题的频率，绘制散点图

发现：

每个小组都会出现需求、设计、开发、接口、部署的邮件主题

猜想：

hightech有三条业务线，为三个组，其下小组又负责不同功能模块



计划

1. 联动部门结构图与部门邮件主题 => 小组职能更加明确?
2. 将邮件主题分类
 1. 一级类别：研发、人力资源、财务、综合、研发-监控
 2. 二级类别：研发-需求、研发-开发...