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）

答　卷

参赛队名称：中国科学院信息工程研究所-陈明毅-挑战 1

团队成员：　陈明毅，中国科学院信息工程研究所，chenmingyi@iie.ac.cn，队长

蔡真真，中国科学院信息工程研究所，caizhenzhen@iie.ac.cn

韩瑶鹏，中国科学院信息工程研究所，15152108971@163.com，

蹇诗婕，北京科技大学，17888838363@163.com

田甜，中国科学院信息工程研究所，tiantian@iie.ac.cn，指导老师

团队成员是否与报名表一致（是或否）：是

是否学生队（是或否）：是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：Echarts，Tableau，vue，MySQL

共计耗费时间（人天）： 60 人天

本次比赛结束后，我们是否可以在网络上公布该答卷与视频（是或否）：是

挑战 1.1：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。

本题思路从邮件入手，利用分词聚类算法，基于每位员工会向上一级领导提交月报的假设，筛选出邮件标题为“总结”、“近期工作总结”、“工作汇报”、“月度总结”、“项目汇报”等等词汇的邮件数据，对每一条邮件记录的 Sender 和 Receiver 做关联，形成如图 1-1 所示的工作汇报结构图。点击每个节点，显示该节点的子节点和二级子节点的邮件标题词云图，通过常识认知，根据词云内容判断所属部门。例如，点击 1059 这个节点，可看到其麾下组长和普通职工的邮件标题词云图，从而判断该部门为研发部门。

图 1-1 工作汇报结构图及邮件词云图

对每个部门的词云内容进行判断，得出部门结构图如图 1-2 所示。HighTech 公司共有 299 名员工，总经理为工号 1067，统领财务、人力资源和研发三个部门，对应的员工人数分别是：24、18、256。财务部长工号为 1041，人力资源部长工号为 1013，研发部门划分为 3 个部门，由工号 1007、1059、1068 三位部长分别管理。研发 1、2、3 又分别划分为 9、11、7 个小组，每个小组设有组长。

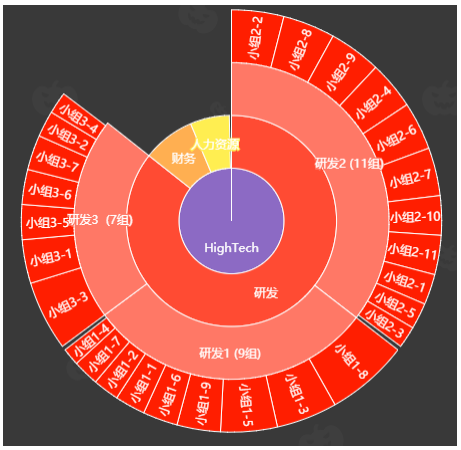

图 1-2 公司部门结构图

通过上述分析，得出员工的组织结构图如图 1-3 所示。点击每个部门或小组，可以查看该部门或该小组的职工人数，点击每位员工，可以查看员工所处职位。

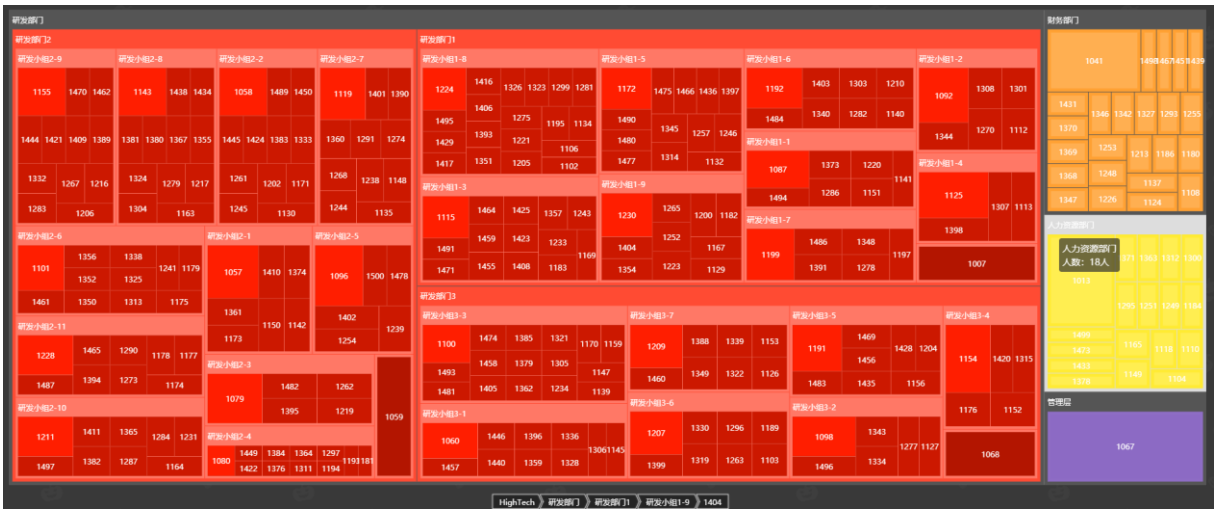

图 1-3 公司内部员工组织结构图

挑战 1.2：分析该公司员工的日常工作行为，按部门总结并展示员工的正常工作模式。

一. 任务下达以及工作汇报

图 2-1 公司各部门员工及管理邮件主题词云图

如图 2-1 是公司各部门员工以及管理的邮件主题词云图，结合 Tableau，对每个主题的邮件发送时间进行统计，发现并无明显时间规律。由图 2-1 可知，总经理向五个部长下发《年度计划》和《公司发展规划》，五个部门贯彻落实。

1. 财务部门

财务部门主要负责财务报账、资金、税务、成本控制和会计核算，员工每周需向财务部长提交工作汇报，部长每周向总经理提交工作汇报。

2. 人力资源部门

人力资源部门主要负责招聘、面试、考勤管理和绩效考核。部长负责下发工作计划，且参与招聘工作，组员不定期向部长提交《招聘信息总结》，部长每周向总经理进行工作汇报。

3. 研发三大部门

研发部门主要负责项目开发。三位研发部长每周不定期组织 1-2 次例会，下达本周工作任务，各个组长将其整理成《例会会议纪要》下发给组员。员工向组长提交工作总结，组长向部长提交月报总结，部长向总经理不定期进行工作汇报。各个研发小组每周由组长安排技术分享。

二. 各部门上下班打卡情况

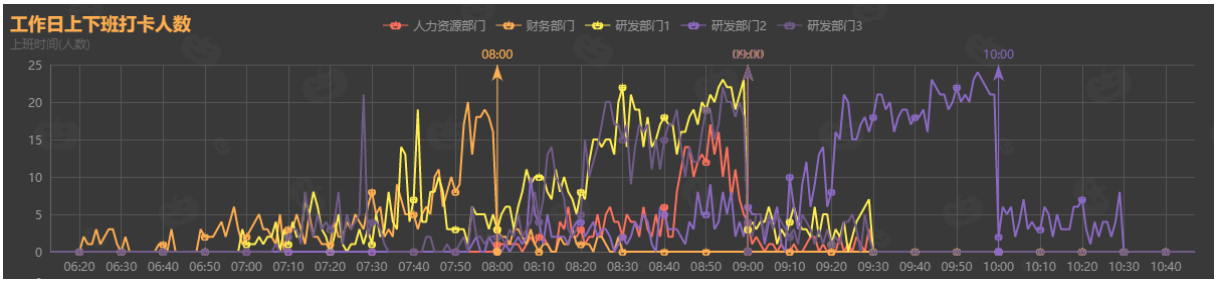

图 2-2 工作日上班打卡情况

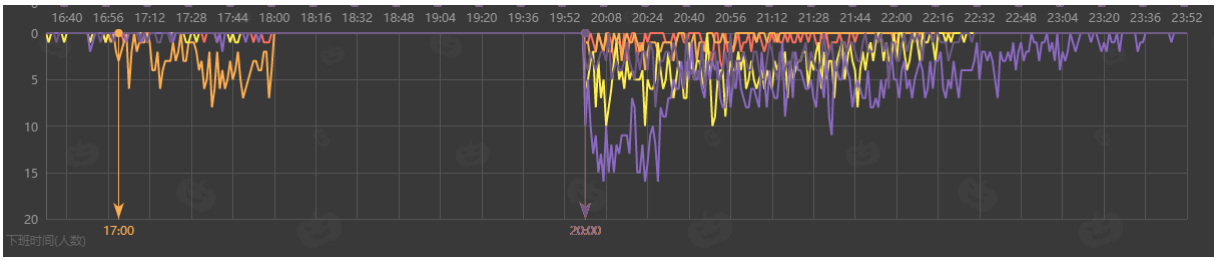

图 2-3 工作日下班打卡情况

HighTech 公司五个部门的上下班打卡情况如图 2-3、2-3 所示，可以通过梯度下降和梯度上升的情况判断每个部门的上下班规定时间，当上班时间出现断崖式下跌和下班时间出现跃进式提升，结合 kaoqin@hightect.com 邮箱第二个工作日会给前一天迟到或者早退的员工发《迟到》、《早退》邮件进行二次验证，得出五个部门的上下班时间如下表所示：

表 2-1 五个部门的上下班时间

部门	人力资源部门	财务部门	研发 1 部门	研发 2 部门	研发 3 部门
上班时间	9:00	8:00	9:00	10:00	9:00
下班时间	20:00	17:00	20:00	20:00	20:00

研发部门 2 上班时间较另外两个研发部门晚一个小时，但是该部门下班打卡曲线较为滞后，时常加班到 22 点之后。

三、正常工作模式

对五个部门每日的流量访问情况进行统计，每十分钟为一个单位，如图 2-4 所示，可发现每个工作日的流量曲线波动情况基本一致，出现三个谷值，上班时间出现谷值意味着有大量人员暂停工作。根据前面提供的信息，可以判断这三个谷值可能为例会、午休、技术分享时间。具体情况如表 2-2 所示。

表 2-2 五个部门工作时间安排

部门	人力资源部门	财务部门	研发 1 部门	研发 2 部门	研发 3 部门
9:30-10:00	例会	例会			
10:20-10:50			例会	例会	例会
12:30-13:00	午休	午休	午休	午休	午休
13:20-13:50	午休	午休	技术分享	技术分享	技术分享

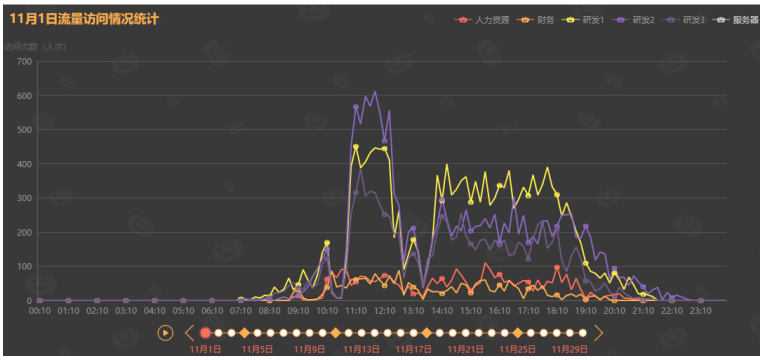

图 2-4 工作日下班打卡情况

四、加班情况

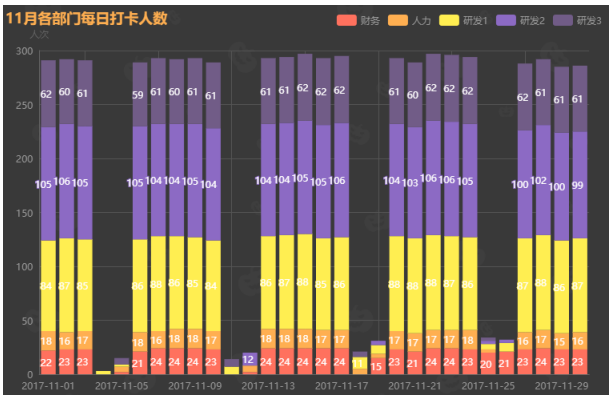

图 2-5 各部门每日打卡人数

对该公司每日打卡人数进行统计，如图 2-5 所示，可得该公司实行一周双休制，周末无特殊情况不加班。但在 11 月 15 日、20 日、21 日财务部门组织加班。

五、服务器资产模式

如图 2-6 所示，通过三个研发部门员工对服务器的访问情况，得出绝大部分服务器是各个研发组混合使用，并且没有专门的用途，同时运行多个服务（尤其是多个类型的数据库）。个别服务器是专门服务于某

个研发部门。同时，HighTech 公司在公网上部署有多台服务器（或是合作伙伴服务器）。部分研发人员也有自己的服务器，涉及境内外数百台服务器。

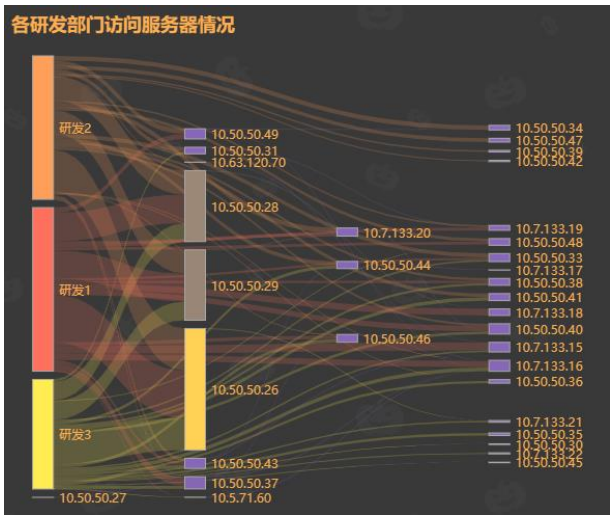

图 2-6 各部门每日打卡人数

挑战 1.3：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。

一、1281、1376、1487 三人同一天离职事件

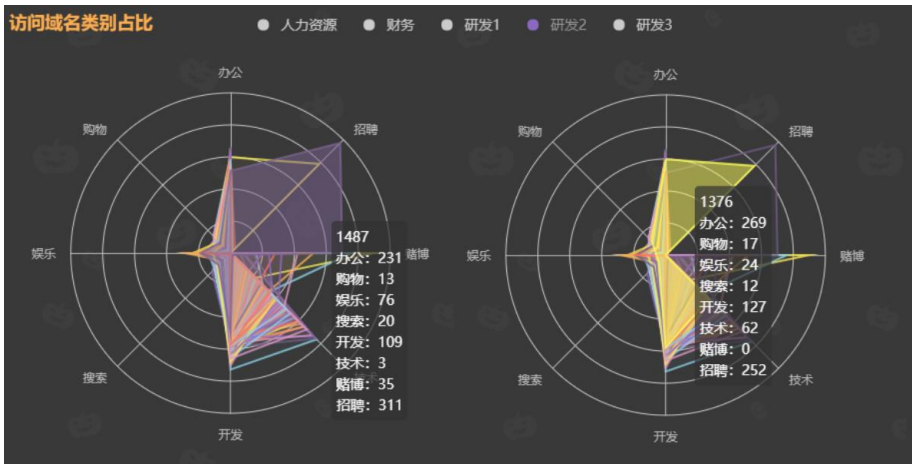

图 3-1 各部门员工访问域名类别雷达图

对各部门员工访问域名类别进行雷达图分析，发现工号为 1376、1487 两名员工在招聘区域的面积非常大，代表其大量浏览招聘网站。查询邮件记录，发现 1281、1376、1487 三名分别来自研发 1-8、2-4、2-11 的普通员工于 11 月 27 日下午向 hr@hightech.com 邮箱提交【辞职信】，11 月 28 日上午分别获得所在组组长和所在部门部长批准。同一天辞职，实属异常，是否包含被辞退的可能。

二、1487 盗用组长账户事件

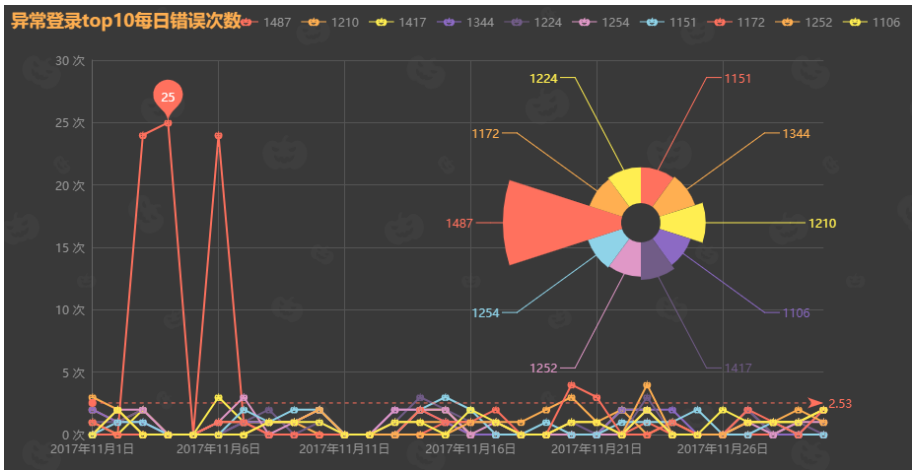

图 3-2 异常登录 Top10 每日错误次数

选取 Login 表中登录状态错误次数 Top10 的员工进行可视分析，可以发现工号为 1487 的员工出现了峰值，其在 11 月 3 日、4 日、6 日这三天频繁登录错误。对该名员工进行个人分析，如图 3-3 所示，发现该名员工就是第一个异常事件的离职员工之一，他大量访问招聘网站，于 28 日正式离职。该员工在 4 日、25 日有登录服务器情况，但当天并无打卡记录。

图 3-3 对 1487 进行个人分析

由图 3-4 所示，通过平行坐标图，可以发现员工 1487 所使用的主机 IP 连线了 4 个用户，并且可以通过颜色看出，其供成功登录 2 个用户。除开自身，其在 11 月 3 日、4 日分别使用 2-4 组长 1080、2-10 组长 1211 的账号登录 IP 为 10.50.50.44 的服务器，由于密码错误，均登录失败。在 11 月 6 日尝试使用 2-11 组长 1228 的账号登录，最终于 22 点左右猜中密码，登录成功。其后，又在 16 日 20 点、24 日 12 点半左右成功登录。

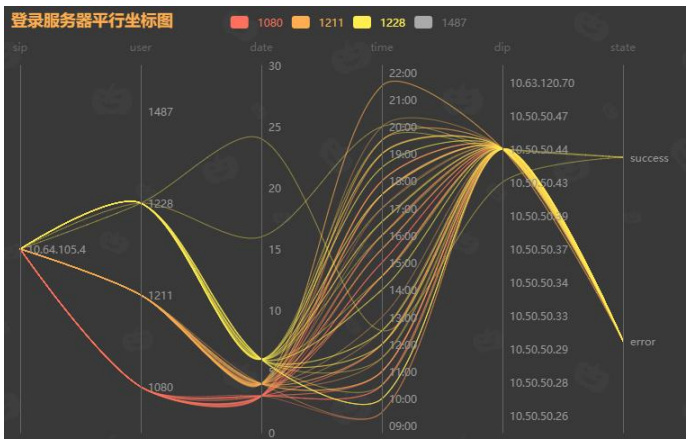

图 3-4 1487 盗用账号平行坐标图

而查看 1487 的流量折现图，发现其在 16 日 20 点盗用组长 1228 的账号登录 44 服务器后，出现了流量峰值，代表其使用 ssh 协议下载了大量数据。

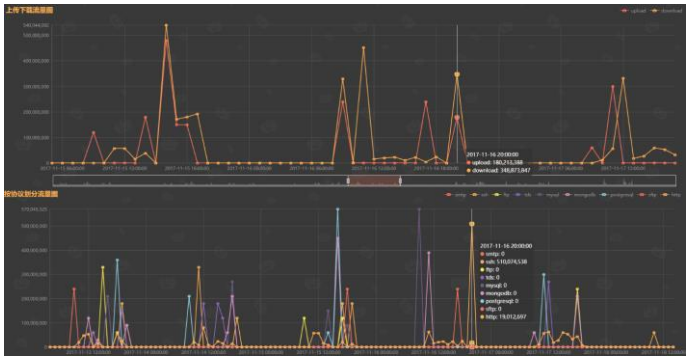

图 3-5 员工 1487 流量图分析

三、1487 等 5 人团伙泄密事件

分析 1487 在 24 日 12 点 44 分盗用组长 1228 账户的异常登录行为，发现其通过 10.50.50.43 做跳板登录 10.50.50.44 服务器，并向境外服务器 13.250.177.223 上载了 572M 数据。同样利用平行坐标图分析 13.250.177.223 服务器相关的溯源连接记录。发现仅有五条访问记录，对其进行追踪溯源，发现 1183，1273，1487，1169 和 1151 五人，均经过两次跳转，向该服务器上载数据，并且都进行了大量数据的上载。这种显而易见的跳转隐匿行为具有极高风险，且 1487 利用组长账号做掩饰，更有掩耳盗铃之嫌疑。

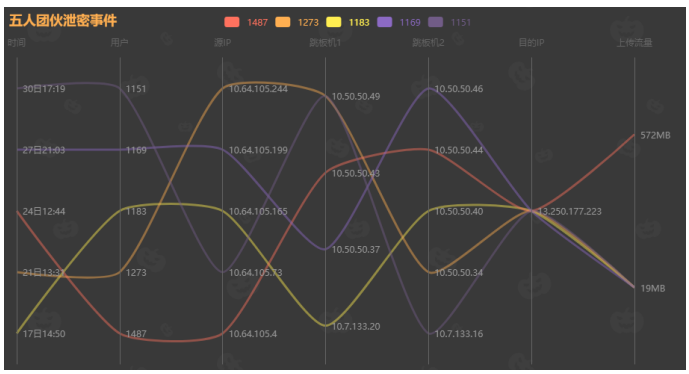

图 3-6 五人团伙通过跳板机向境外服务器上传数据

四、1376 意图破坏数据库，窃取数据事件

分析 alert 告警邮箱的主题词云，发现其在 16 日晚 20 点向员工 1284、1487 发送了数十封数据库异常告警邮件。分析两位员工登录服务器的平行坐标图，发现其均在 20 点 30 分之后登录 10.63.120.70 服务器进行维修。对事发前后 70 服务器的流量的流量进行可视分析，发现在报警发生之前，只有 1376 使用 ssh 协议频繁访问服务器，并下载大量内容，造成了明显的峰值，具有破坏服务器并窃取数据的动机。恰巧，该事件中的员工 1376、1487 均在同一天离职，是否有伙同犯案嫌疑。1487 在 8 点 30 分登录服务器进行维修，是否有打扫现场的嫌疑。

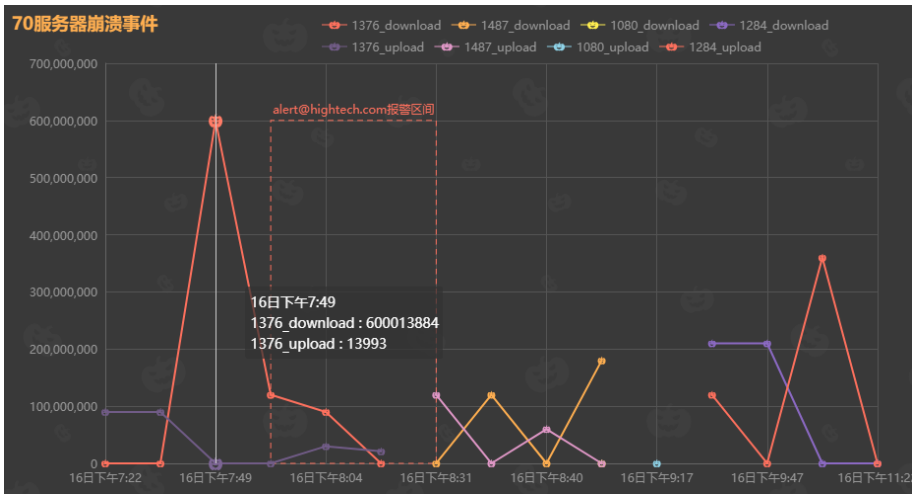

图 3-7 10.63.120.70 服务器崩溃事件

五、员工 1376、1487 等 9 人小团体具有潜逃风险

通过对公共服务邮箱进行分析，发现 hr@hightech.com 每周四早上 9 点半会给全体员工发送主题为《打球啦，欢迎大家参加》邮件，参与的员工会进行回复，通过对每周参与活动的员工进行关联分析，发现 1149、1261、1313、1330、1352、1376、1383、1389、1487 这 9 名员工均只参与前三周的打球活动，极有可能是一个小团体。

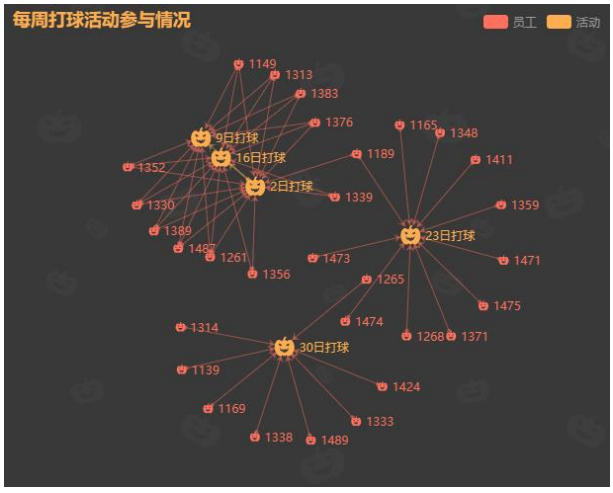

图 3-8 每周打球活动参与员工分析

对该 9 名员工的打卡记录和邮件关联关系进行可视分析，利用日历图可以发现员工 1149(人力资源)、1352(研发 2-6)、1383(研发 2-2)、1389(研发 2-9)该四名员工在 27 日至 30 日四天均旷工（没有工作时长），且无共同出差可能。因此猜测 4 名员工极有可能伙同打球团体中已离职的 1376、1487 进行核心数据窃取并潜逃。建议对剩余三名员工 1261、1313、1330 进行调查。

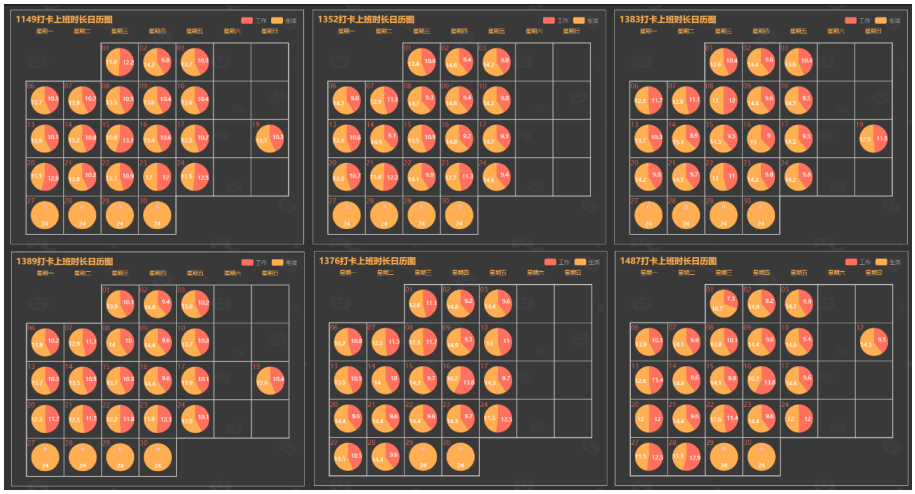

图 3-9　9 人小团体异常员工打卡记录

六、孤立异常事件

6.1 服务器没有按计划访问

服务器 10.50.50.27 在每周六凌晨 3 点固定访问服务器 10.7.133.15，而在月末的 25 日却没有任何访问记录。

6.2 公司过量招聘

该公司仅有 299 名员工，但于 11 月发了 149 封 Offer 和 149 封录用通知邮件，招聘人数与公司规模严重不符。

6.3　研发组长充当猎头

该公司 20 位研发组长收到累计 257 封来自猎聘网的候选人邮件，职位申请包括政府事务经理、销售经理、组织发展总监等，研发组长均注册猎头账号招人，且申请职位与研发无关，实属异常。