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：　　东北师范大学-徐劭斌-挑战 1

团队成员：　徐劭斌，东北师范大学，2233935216@qq.com，队长

任珂，东北师范大学，104431931@qq.com

邹亮，东北师范大学，tail107@nenu.edu.cn

田良，长春工业大学，1074491684@qq.com

张慧杰，东北师范大学，zhanghj167@nenu.edu.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，D3，MyEclipse ，MySQL，Spyder，Matlab

共计耗费时间（人天）：　30 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

（灰色字为参赛信息填写模板，请参赛者在提交时参照模板填写）

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。
（请将回答尽量控制在 500 字和 5 张图片内）

邮件作为公司人员交流的主要方式之一，存储着员工之间的工作关系、沟通主题等有价值的信息。针对该题，我们基于邮件数据，使用随机森林算法将员工进行部门划分，构建员工间的关系网络，结合信息熵探索部门内部的组织结构。

（1） 部门划分

邮件主题反映了各员工的工作范围，因此它成为判别员工所属部门的重要特征。基于随机森林算法，我们将员工分为财务、人力、研发三个部门，步骤如下：

a. 提取所有数据中频数最大的 90 个主题，为每个员工构建一个 90 维的向量，分别存储该员工收发该主题邮件的频数。

b. 随机选取 40 名员工，基于邮件主题人工赋予其所属部门的标签，构建为训练集，并进行随机森林训练。

c. 利用训练好的随机森林将剩余的 259 名员工分类。

我们发现财务部门有 24 名员工，人力资源部门有 18 员工，研发部门有 257 名员工。

（2） 探索部门内部组织结构

根据员工间收发邮件的关系，我们构建各个部门的节点连接图（如图 1-1）。图中每个节点代表一名员工，节点的半径代表员工在部门内部收发邮件的总数，边的粗细编码两名员工邮件往来数量。另外，我们使用信息熵衡量每个员工自我中心网络的混乱程度，即信息熵越大，该员工有更多的联系伙伴；反之，该员工只与个别人员关系紧密。我们使用信息熵编码节点的颜色，从绿到橙映射信息熵从小到大。

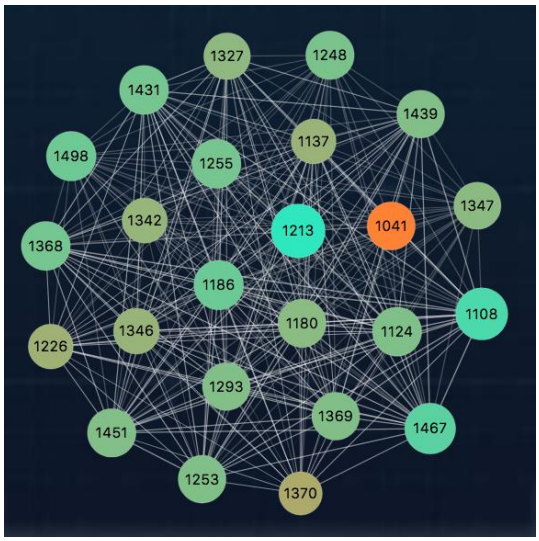

图 1-1 财务部节点连接图

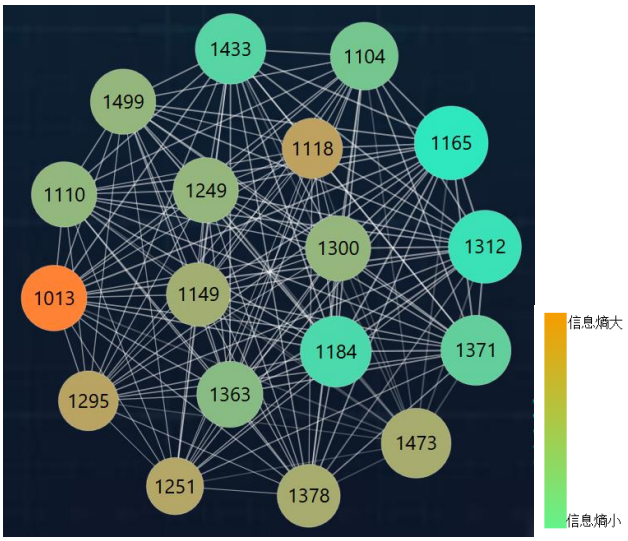

图 1-2 人力资源部节点连接图

从图 1-1 及图 1-2 中可以看出，财务部和人力资源部内部邮件往来密切，没有出现分堆现象。由于领导管理整个部门，与所有员工均有大量收发邮件行为，信息熵比普通员工更大。图 1-2 中 1013 号员工的节点为橙色，其信息熵最大，当我们使用鼠标悬停该节点时，与之有关联的节点和边被高亮出来（图 1-3-a），即展示该员工的自我中心网络，发现他与财务部所有员工均联系密切。我们统计他的收发邮件主题，以文字云的形式展示出来（图 1-3-b）。他收到的邮件中主题为“工作汇报”的最多，而发件主题中“年度工作目标”最多，由此我们可以判定 1013 号员工为人力资源部领导。同理我们探索 1041 号员工与财务部人员的关系及收发邮件主题，判定他为财务部门领导。

图 1-3　1013 号员工信息概览

如图 1-4 所示，研发部门员工聚集成明显的簇，即可分为相对独立的 27 个部门单元。我们也可发现其组织结构，根节点 1067 号为最高层领导，1007、1068、1059 号为二级领导，他们分别管理 9、7、11 个部门单元。而每个部门单元与外界联系的为其单元的领导，例如图中 1087 号员工为一个单元的领导。

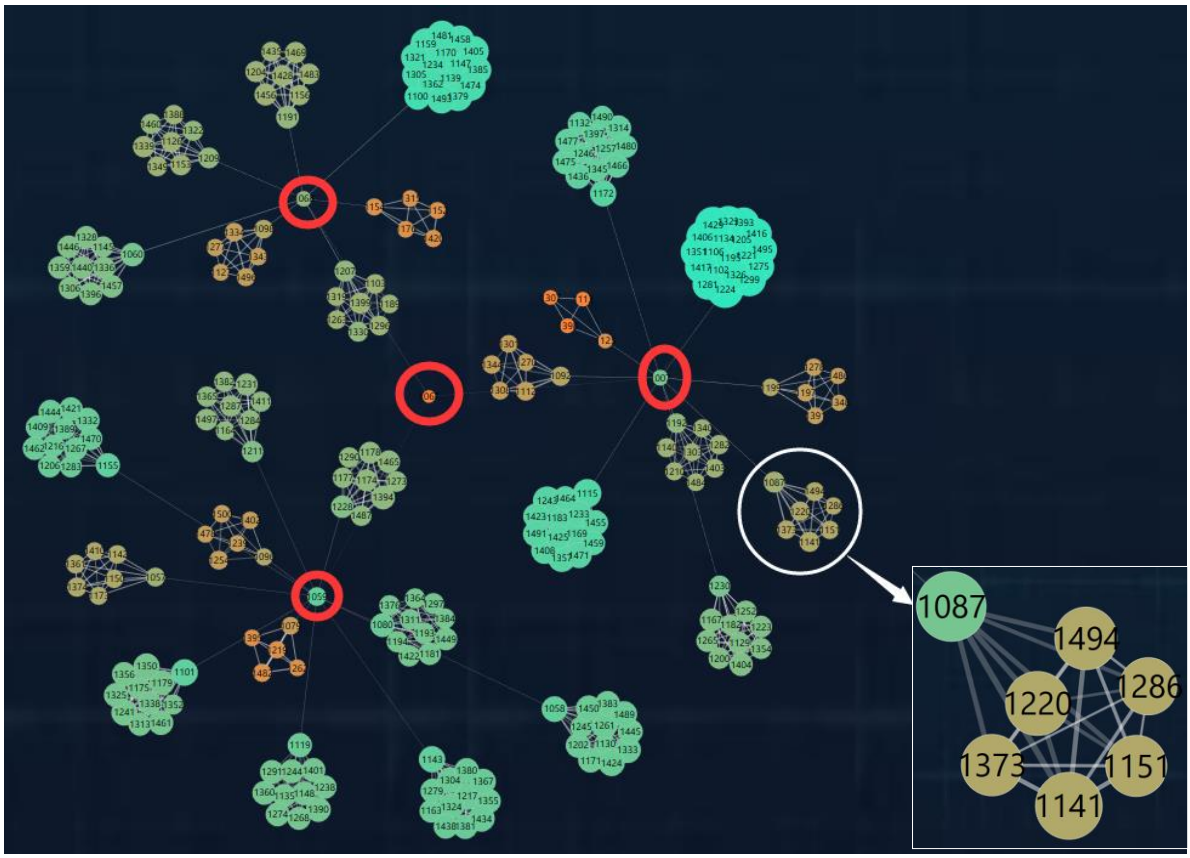

图 1-4 研发部节点连接图

（3）　探索公司组织结构

在数据库查询 1041 和 1013 员工向上级发送工作总结邮件的情况，发现他们只给 1067 号员工发过主题为“工作汇报”的邮件。因此三个部门组织关系便可整合在一起得到整个公司的组织结构图（图 1-5）。1067 号为最高领导，1041 为财务部领导，1013 为人力部领导，1007、1059、1068 为研发部领导。

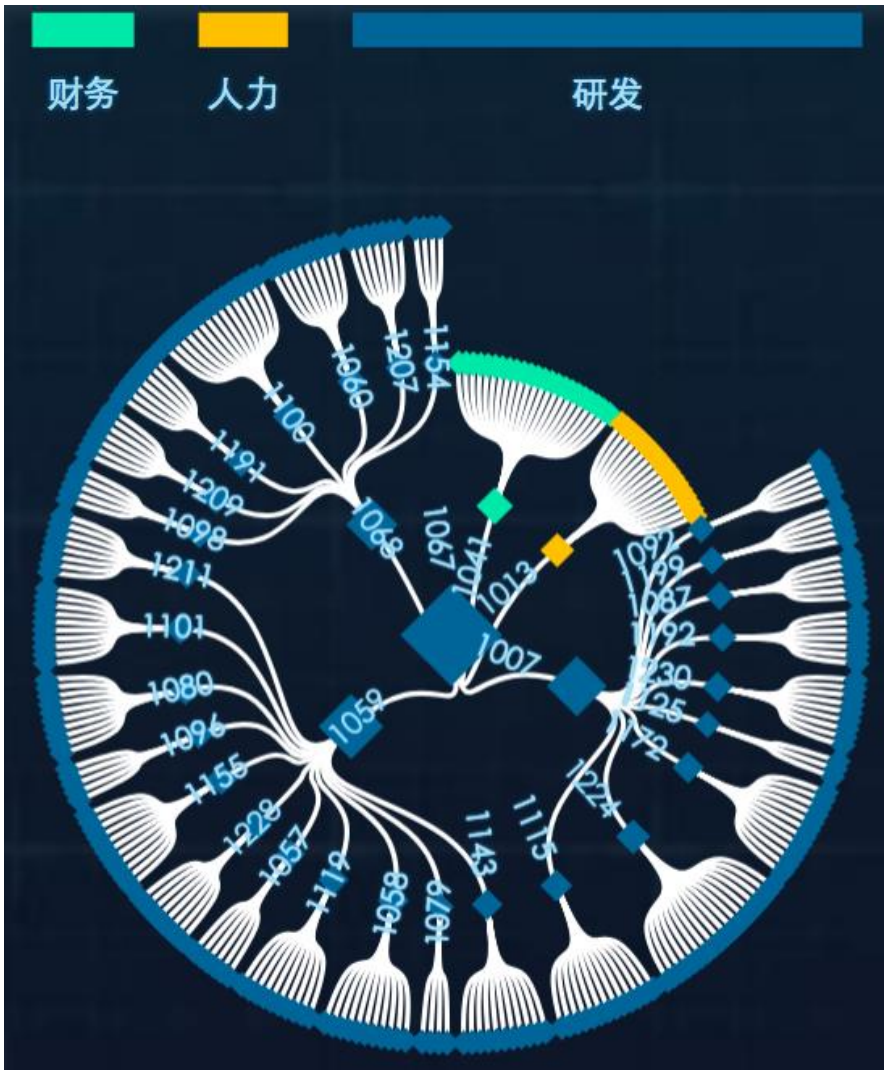

图 1-5 公司组织结构图

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。（请将回答尽量控制在 1000 个字和 8 张图片内）

我们将从工作时间、考勤情况、登陆日志及上下行流量、工作内容几个方面探索各部门的正常工作模式。

（1）财务部门：

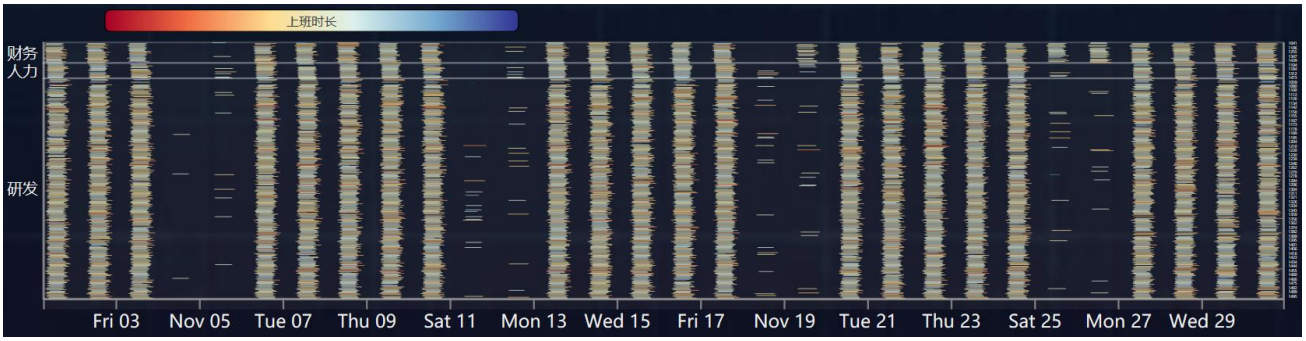

图 2-1 员工考勤热力概览图

图 2-2 财务部门员工工作行为概览

工作时间：图 2-1 展示了全公司各员工的上班情况，每个员工一行，横轴为时间，若员工上班则为这个时间段填色，颜色映射该天的上班时长。可以看出财务部在 11 月 19 日（周日），25 日（周六），26 日（周日）出现超过 60%比例的大面积加班情况。我们推测月底财务部门需要进行账目整理，因此导致这一现象。图 2-2-a 及图 2-2-b 展示了财务部员工上下班时间的分布情况，横轴为时间，纵轴为频率。员工从 7:00 开始陆续上班，在 7:50 到 8:00 间达到峰值，之后大幅减少，因此财务部门正常上班时间为 8:00。同时员工从 17:00 开始下班，在 17:30-18:30 间达到高峰，可以推测该部门正常下班时间为 17:00。总之，**财务部门的工作时间为 8:00-17:00，平日存在短时间的加班现象，月底加班严重。**

考勤情况：图 2-2-c 展示了该部门的 11 月份的员工考勤异常情况，包括：迟到早退、旷工、辞职。图中每个柱形代表一名员工，柱形高度代表异常次数，水波图的比例代表出现异常情况的员工占部门总员工的比例。**财务部门员工迟到早退现象严重，高达 75%。**

登陆日志及上下行流量情况：图 2-2-d 展示了财务部门的登陆日志及上下行流量数据，我们将各个属性分段后，统计重复记录的数量，绘制平行坐标图，一条线代表一条记录，最后一个轴标示该记录的数量。图中只有 2 条线，说明**财务部门情况单一，不涉及 login 登录日志数据（TCPLOG 日志主要是 http 协议的网页访问行为和 smtp 协议的邮件收发行为），下行流量远大于上行流量。**

工作内容：图 2-2-e 展示财务部门频数最多的 10 个邮件主题，主要为财务报账、资金、会计核算等关键词，推断出**财务部门的工作内容为财务账目的整理。**

（2）人力资源部门：

工作时间：从图 2-1 中发现人力部门每周末有 5 人左右加班。从图 2-3-a 中发现，员工从 8:00 开始陆续上班，在 8:50 到 9:00 区间达到峰值，之后骤减，推测财务部门正常上班时间为 9:00。从图 2-3-b 中发现，18:00 过后下班人数激增，推测此时为人力资源部下班时间。总之，**人力资源部门的工作时间为 9:00-18:00，存在少量晚上加班及周末加班的现象。**

考勤情况：如图 2-3-c 所示，**人力资源部门员工迟到早退现象严重，旷工现象较其他部门严重。**

登陆日志及上下行流量情况：如图 2-3-d 所示，与财务部门类似，人力资源部门员工**不涉及 login 登录日志数据，下行流量远大于上行流量。**

工作内容：如图 2-3-e 所示，邮件主题主要为公司简介、复试通知、offer 等关键词，则**工作内容主要是进行对外宣传和招纳新员工。**

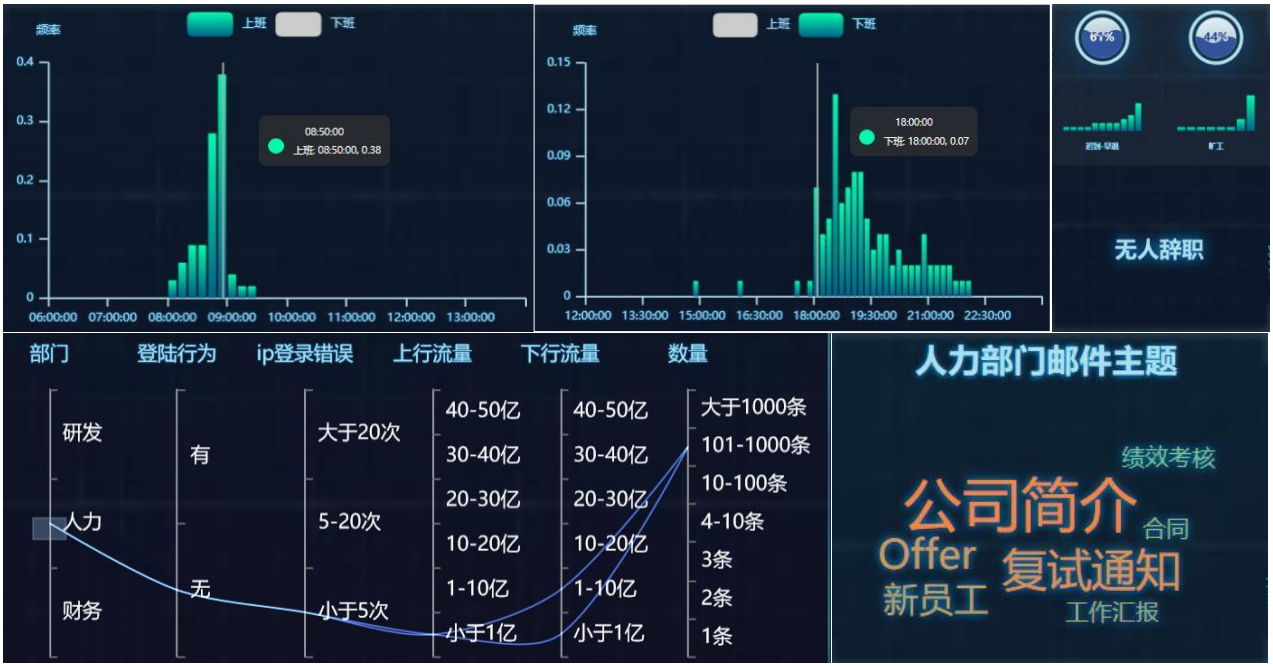

图 2-3 人力资源部门员工工作行为概览

（3）研发部：

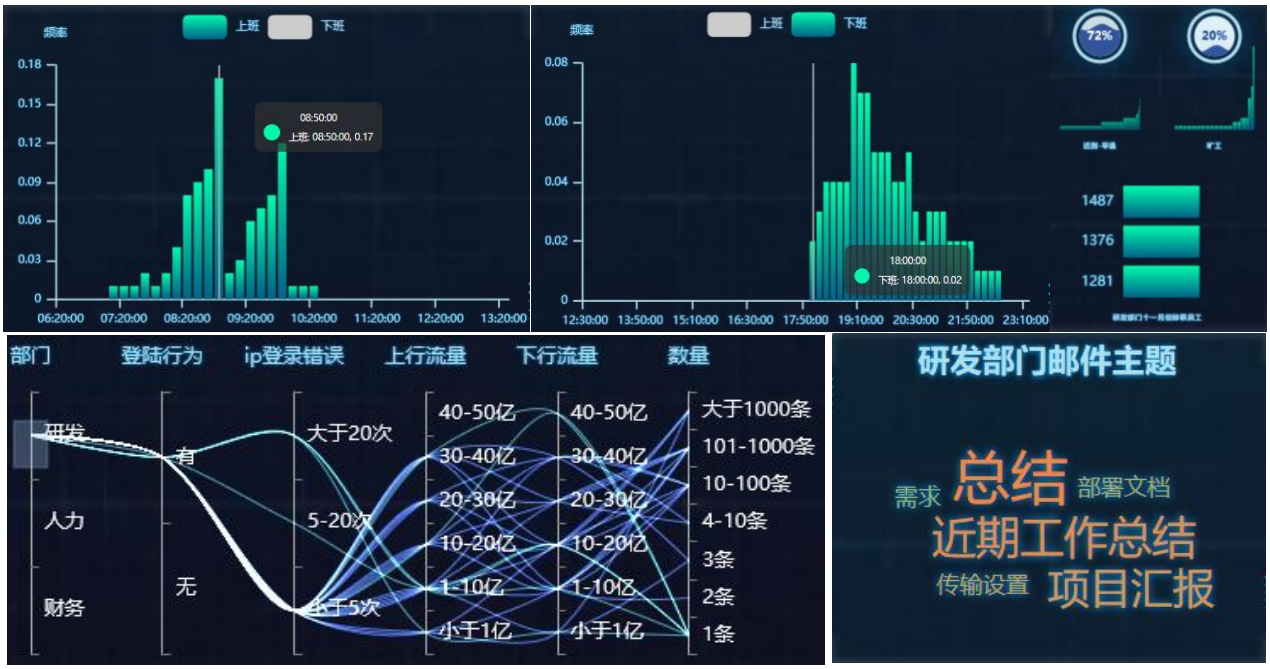

图 2-4 研发部门员工工作行为概览

工作时间：如图 2-4-a 及 2-4-b 所示，研发部门上班时间普遍较晚，存在两个高峰期 8:50-9:00 和 9:50-10:00，20:00-23:40 下班的人数人占有较大比例。研发部门规模较大，我们推测下属的三大研发部门有不同的上下班时间。如图 2-5-c 所示，刷选 1059 号领导的研发一部（包含 9 个部门单元），发现上班时间分布为单峰，在 10:00 后上班人数骤减，在 19:00 后研发一部的员工开始下班，所以研发一部的正常工作时间为 10:00-19:00，晚上加班严重。同样的方法，对 1007 号员工领导的研发二部（图 2-5-b）和 1068 领导研发三部（图 2-5-c）进行刷选，观察其部门三十天上下班的分布，发现这

两个部门比较相似，正常工作时间均为 9:00-18:00，晚上加班现象同样严重。总之，**研发部门存在两个上班时间段，即 9:00-18:00 和 10:00-19:00，平日夜晚加班情况严重。**

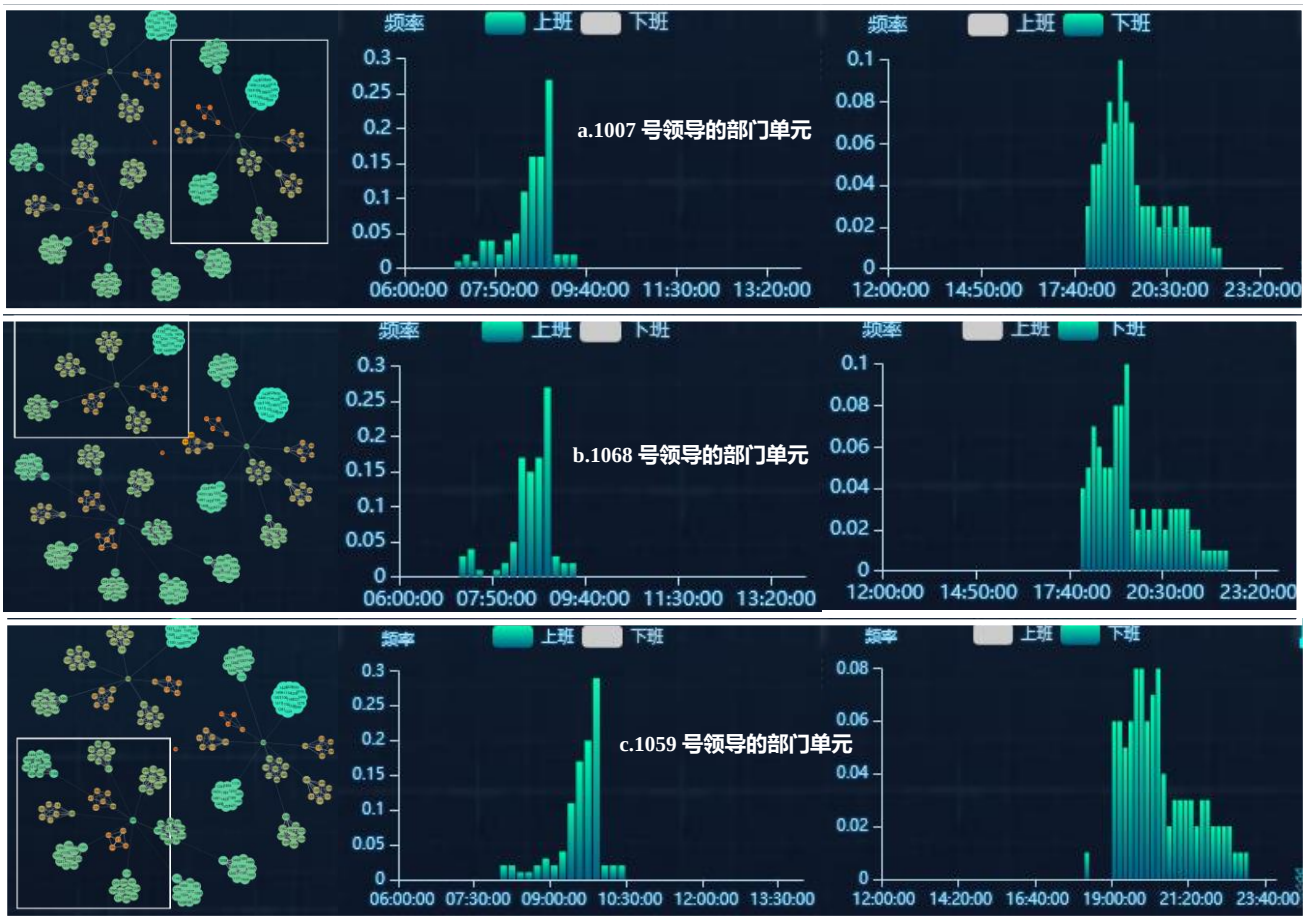

图 2-5 研发部门下三部门员工工作行为概览

考勤情况：如图 2-4-c 所示，研发部门 11 月有 1281、1376、1487 号员工辞职，而财务部和人力资源部门无辞职员工。**研发部员工存在跳槽行为**，公司需要完善运行机制来减少因为员工离职对业务产生的消极影响。

登陆日志及上下行流量情况：如图 2-4-d 所示，我们刷取数量大于 100 条的记录探索研发部门员工的日常行为，发现该部门人员有 ftp，mongodb 等 7 个协议的 login 登录记录，登陆错误较少，**上下行流量不存在明显的大小关系。**

工作内容：如图 2-4-e 所示，邮件主题多为总结、项目汇报、项目分析、传输设置等关键词，则**工作内容主要是进行产品的研发。**

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。（请将回答尽量限制在 1500 个字和 10 张图片内）

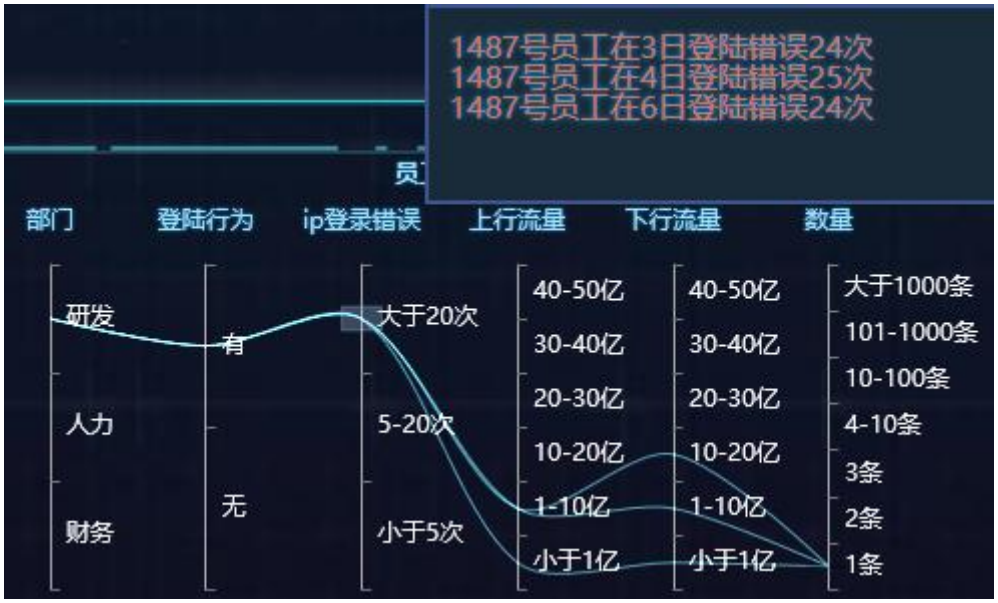

图 3-1 登陆日志平行坐标图（异常登陆错误）

（1）登陆错误：如图 3-1 所示，我们在展示登陆日志及上下行流量情况的平行坐标图中刷取数量为 1 的异常记录，发现一些 ip 登陆错误大于 20 的记录，继续筛选这些数据，发现 **1487 号员工的 ip 存在大量登陆错误行为**。

图 3-2 各协议登陆错误径向堆叠图

（2）越权行为：为了探索 1487 号员工 ip 出现登陆错误的具体情况，我们使用径向堆叠图 3-2 展示他使用自己账号时各协议的登陆情况，半径轴代表 login 相关的 7 种协议，径向轴代表数量。从 3-1 中已知，1487 号员工使用自己 ip 登陆失败的数量大于 20 次，而 3-2 中自己账号登陆失败的数量远不及此，这说明该员工曾使用自己的主机尝试登陆他人账号，并出现了大量登陆错误的现象。公司为了保护内部机密，通常对不同的员工的账户设定的不同的权限，而 1487 号员工这种登陆他人账户的行为是

典型的越权行为。通过交互操作调出 1487 号越权操作的详细记录（图 3-3），发现这些行为多达 73 次，他曾在 3、4、6 日通过 ssh 协议尝试登录 1211、1080、1228 号员工账户，但一直登陆失败。直到 2017-11-06 19:42:57 他成功登录了 1228 号账户访问并 ip 为 10.50.50.44 的服务器，之后 16、24 日他再次登录 1228 账户进行操作。总之，**1487 号员工多次利用自己的主机，越权登陆 1228 号账户访问目的 ip 为 10.50.50.44 的服务器。**

proto	dip	dport	sip	sport	state	time	user
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 16:36:33.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 17:32:00.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 17:32:43.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 18:24:58.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 18:30:01.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 19:30:52.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 19:46:05.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 19:57:21.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 20:10:59.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 21:43:43.	1211
ssh	10.50.50.44	22	10.64.105.4	49200	error	2017-11-04 21:46:28.	1211
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 10:07:44.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 10:17:49.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 10:21:58.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 10:27:00.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 10:37:59.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 11:35:17.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 12:11:20.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 12:30:49.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 12:41:51.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 12:50:15.	1228
ssh	10.50.50.44	22	10.64.105.4	49197	error	2017-11-06 14:09:27.	1228

图 3-3 1487 号员工越权操作

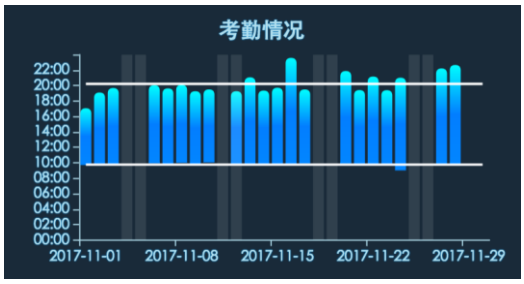

图 3-4 1487 号员工上下班柱形图

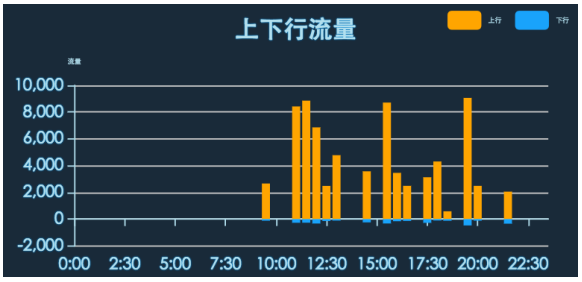

图 3-5 1487 号员工 4 日上下行流量情况

（3）上班打卡异常： 图 3-4 展示了 1487 号员工的上下班情况，横轴为天，纵轴为时刻，柱形的位置和高度代表该天的上班时长，上下两条横线分别代表平均上班时间和平均下班时间。通过上一步已知 4 日该员工曾尝试登陆他人账户，而图 3-4 显示 4 日他并没有打卡上班。图 3-5 展示了员工 tcplog 日志中 4 日的上下行流量情况，我们发现 1487 号员工在 4 日上行流量远大于下行流量，这不符合研发部员工上下行流量属同数量级的常规模式。并且其自身账户没有任何 login 日志的登录记录，由此我们可以推断上行流量全部由尝试登录他人账号产生，而登陆失败导致下行流量远小于上行流量。通过交互点击考勤图中的缺席日期，我们发现 **1487 号员工曾在 4、11、18、19、25、26 日多个周末，隐瞒上班记录。**

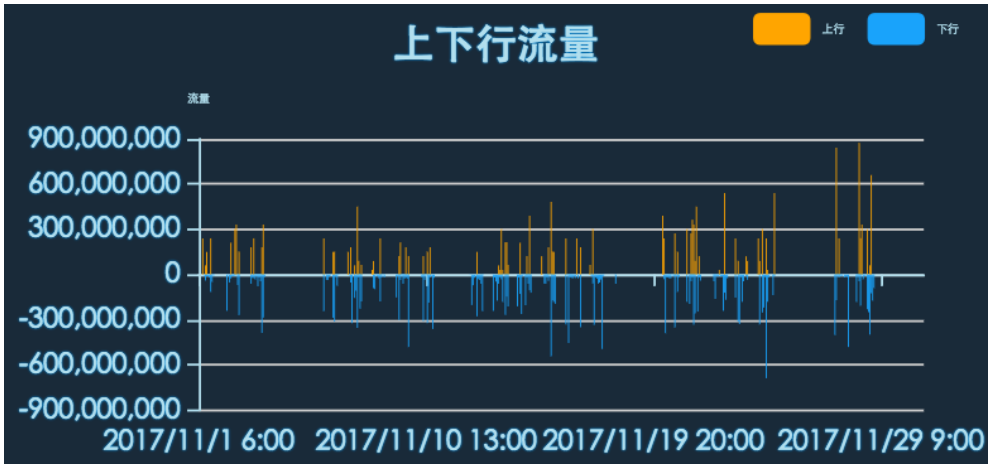

图 3-6 1487 号员工 11 月上下行流量情况

（4）离职前流量异常：在分析部门常规模式时，我们曾利用图 2-4-c 得知研发部门 11 月份有 3 名员工辞职。而 1487 号恰巧为其中一名，他曾在 27 日发出辞职邮件并于 29 日离职。但从图 3-6 中我们可以看出，**1487 号员工在离职前两天 27、28 日的上下行流量突然增大，对服务器进行大量的操作**。

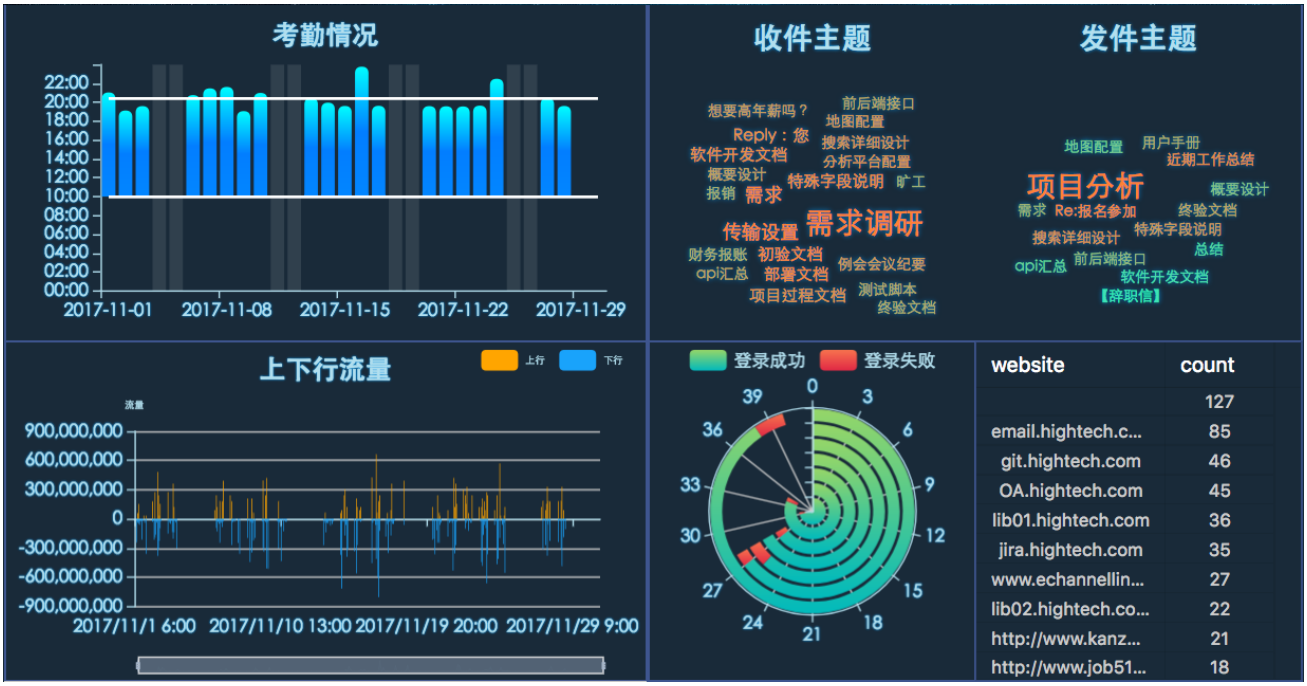

图 3-7 1376 号员工信息概览

（5）同时辞职：查看三个辞职的员工的具体辞职时间，发现 1281、1376 和 1487 号曾在同时 5 个小时内向上级领导发送辞职邮件，并于 29 日同时离职。这一同时辞职事件让我们有理由怀疑 1281 和 1376 号与 1487 号员工的异常行为有牵连。我们以同样的方式探索 1376 号员工的日常行为（图 3-7），发现他曾在 11、12、18、25、26 日无上班打卡记录，却存在上下行流量记录，并且在 25 日他曾通过 ssh 协议产生大量上行流量。另外我们发现 1376 号曾大量访问求职网站，为跳槽做准备。这些都说明**三人的同时辞职并非偶然事件，1376 号员工同时存在隐瞒上班记录的行为。**

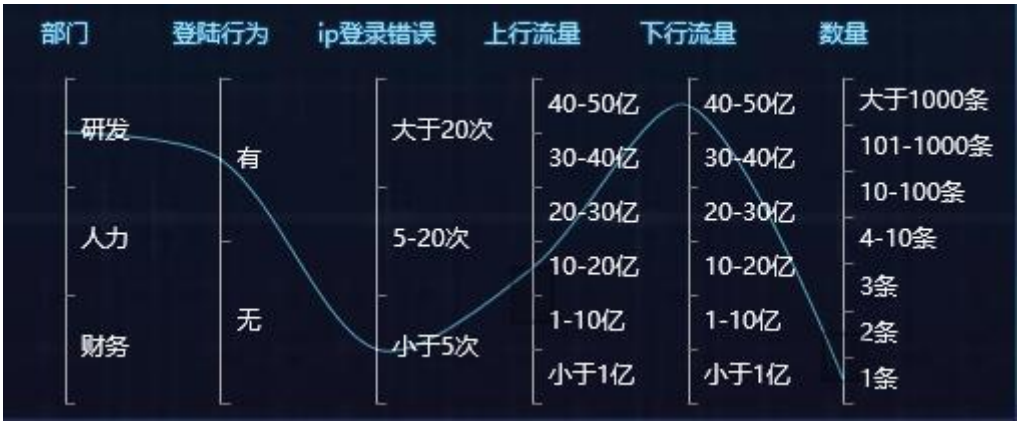

图 3-8 登陆日志平行坐标图（异常流量）

（6）其他异常：通过刷取平行坐标中数量为 1 的记录，我们还发现了一些目前不可解释的异常。例如图 3-8 所示，存在一条下行流量巨大的记录。追溯原始数据，我们发现 **1456 号员工曾在三人离职的 29 日，产生大量下行流量。**

根据所提供的监控数据及多视角可视化设计方案，我们构建了一个完备的威胁情报分析系统 TIVis。我们在进行异常行为探索时共发现上述六件异常事件。可以推断 1487 号员工在辞职前一个月内，选择上班人数较少的周末隐瞒上班行为，大量尝试越权登陆他人账户访问固定的目的 ip，并最终登陆成功窃取信息。1376、1281 号员工与 1487 号员工同时辞职，并且 1376 号存在类似异常行为，我们可以合理推断三人为一个犯罪团伙，其中主犯为 1487 号，主要从犯为 1376 号。另外在三人离职后发生异常的 1456 号员工，应是公司之后的重点观察对象。