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：　上海交通大学-张浩城-挑战 1

团队成员：　张浩城，上海交通大学，jason1120921@qq.com，队长

强志文，上海交通大学，q7853619@sjtu.edu.cn

曹以想，上海交通大学，karen.cao@sjtu.edu.cn

熊　俊，上海交通大学，9763253528023@sjtu.edu.cn

董笑菊，上海交通大学，xjdong@sjtu.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）：D3，Excel，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。

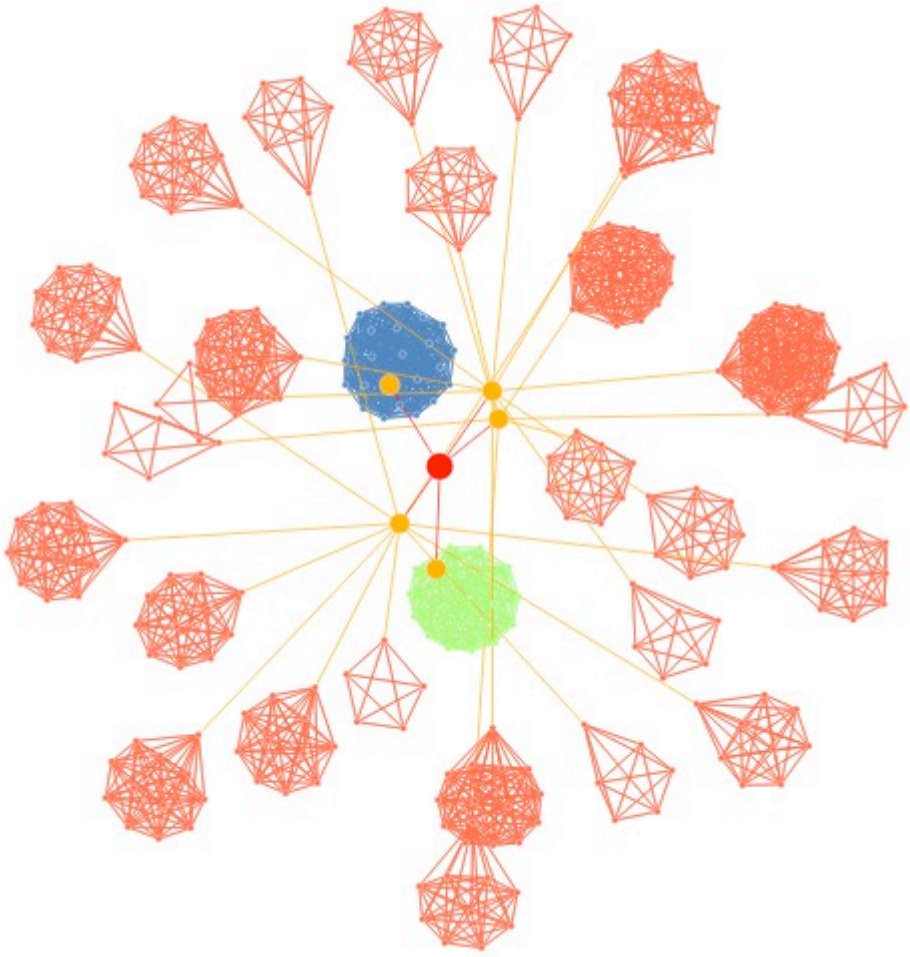

图 1.1

我们用力导向布局算法绘制了公司员工组成图，如图 1.1 所示，我们重点利用了群发邮件的数据，图中每个节点代表一个员工，节点之间的连线代表他们之间有群发邮件的联系。通过分析群发邮件的标题，我们发现公司内部可分为研发，财务和人力三个部门。其中，蓝色表示财务部门，绿色表示人力部门，橙色表示研发部门。红色节点表示总裁，黄色节点表示各部门领导。我们可以发现财务部门和人力部门规模较小，群体内部之间邮件联系非常紧密。两部门各有一个部门领导。负责统筹规划部门事物以及和总裁沟通交流。橙色表示研发部门，我们可以看出研发部门规模较大，分为若干个小群体，群体内部联系十分紧密。研发部门有三个部门领导，负责规划分配部门事物以及和总裁沟通交流。同时研发部门每个群体都有一人和部门领导直接联系，我们推断其为研发部门二级领导。

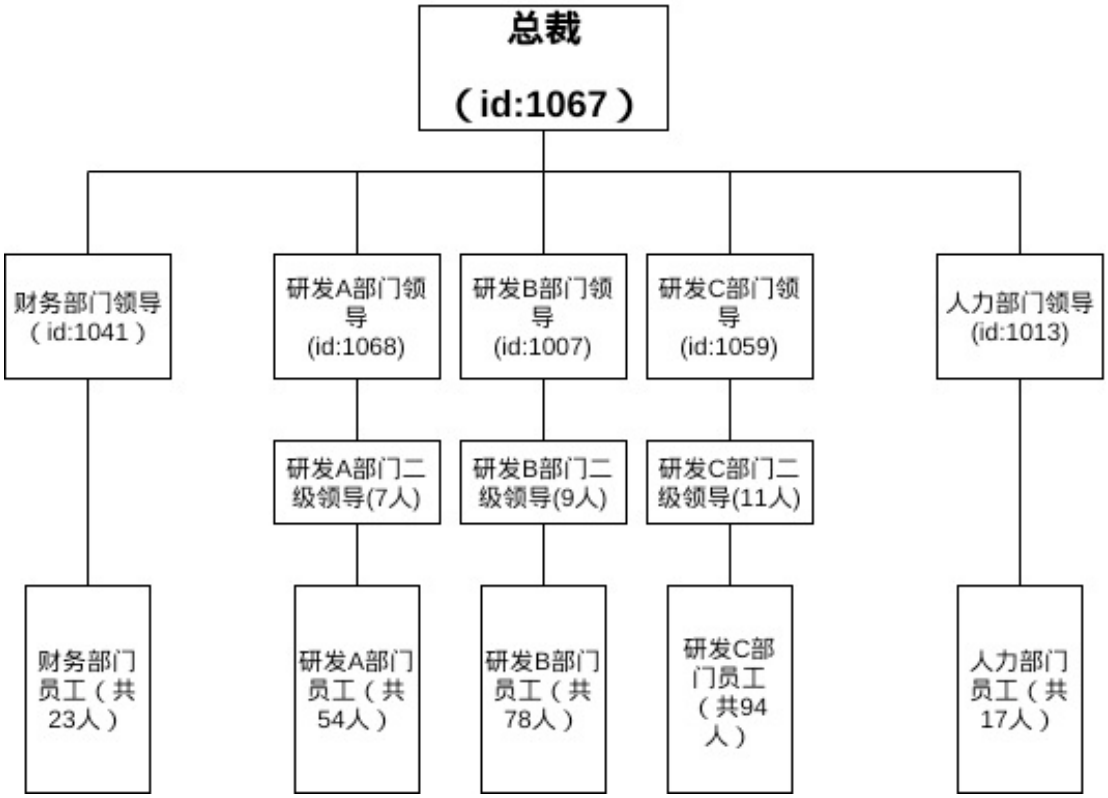

图 1.2　公司员工组织结构图

公司员工的组织结构图如图 1.2 所示，公司一共有 5 位部门领导， 1 位总裁和三位研发部门的二级领导。总裁 id 为 1067，财务部门领导 id 为 1041，该部门员工共有 23 人；研发 A 部门领导 id 为 1068，7 个二级领导 id 分别为 1154，1191，1207，1100，1098，1209，1060。该部门员工共有 61 人;研发 B 部门领导 id 为 1007,9 个二级领导 id 分别为 1087,1115,1230，"1172，1192，1199，1092，1125，1224，该部门员工共有 87 人；研发 C 部门领导 id 为 1059，11 个二级领导 id 分别为 1080，1211，1101，1143，1119，1155，1058，1228，1096，1079，1057，该部门员工共有 105 人；人力部门领导 id 为 1013，该部门员工共有 17 人。

总裁和部门领导直接联系，研发部门再通过部门二级领导将任务和规划下发至各员工。

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。

我们将通过系统依次展示对总裁和财务部门、人力部门、研发部门员工的日常行为分析过程和结论。

总裁：（id=1067）

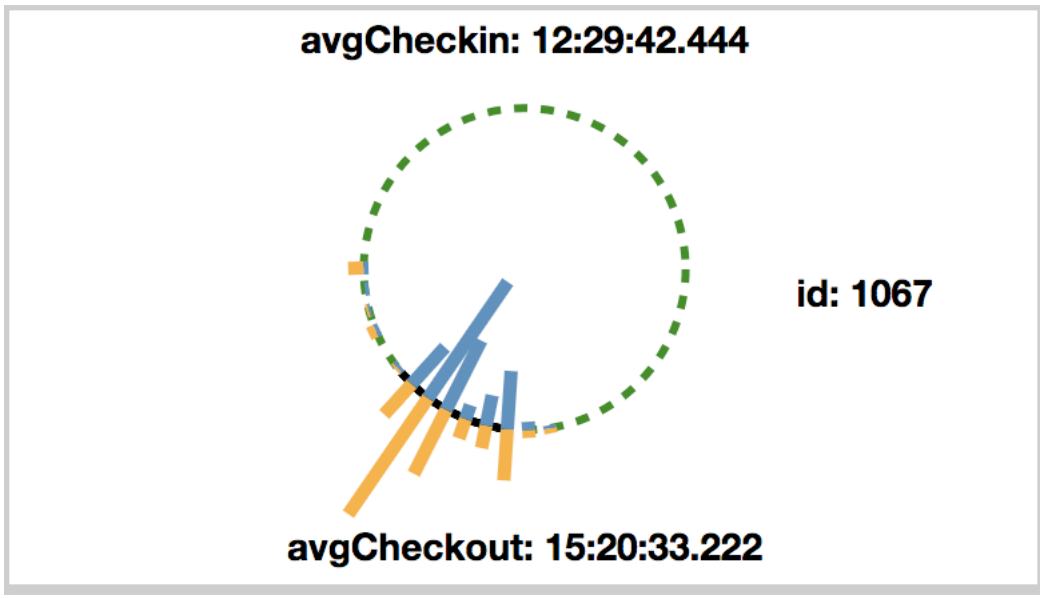

图 1.3

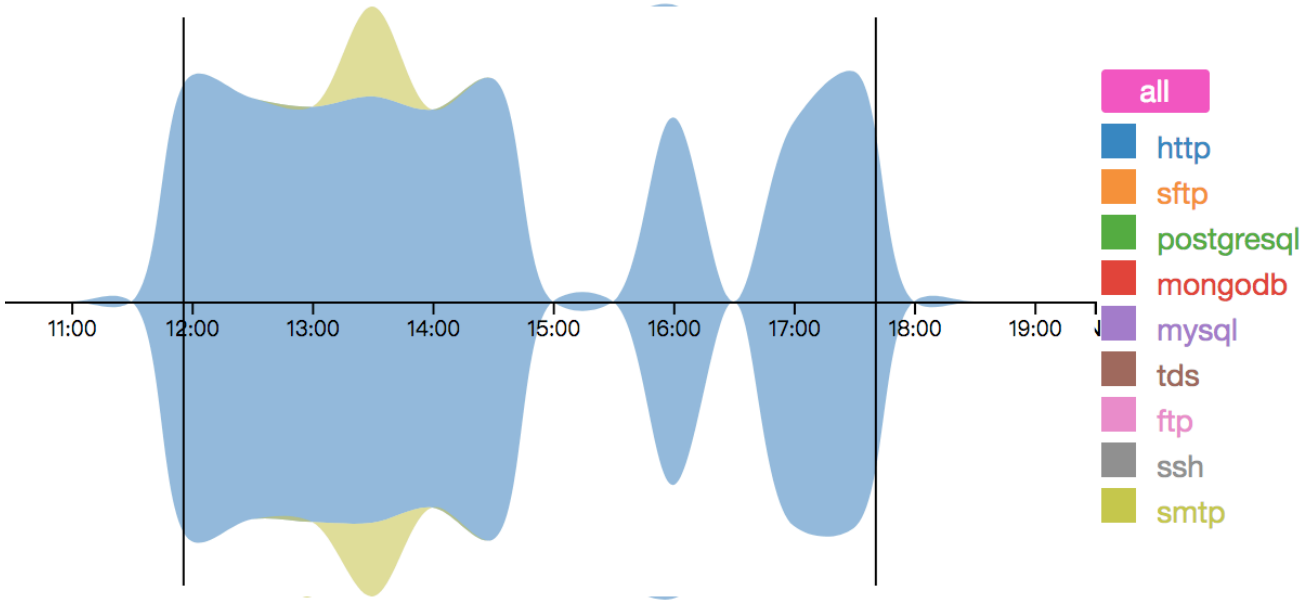

图 1.4

我们计算该 id 平均上下班时间如图 1.3 显示，表明总裁在公司时间并不长且集中在下午，然后用堆叠图以每半小时处理各个协议的数据画出图形如图 1.4，观察可知总裁所用协议基本为 http 和 smtp，这表明，总裁日常工作使用电脑行为中以网上访问行为和发邮件行为为主。

财务部门：

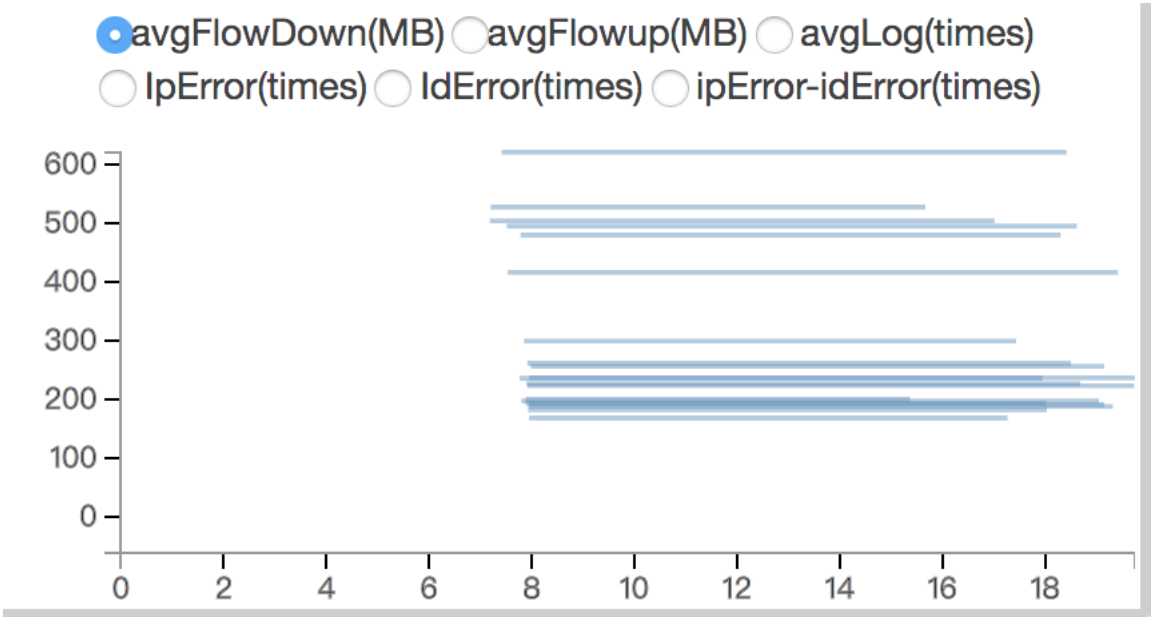

图 1.5

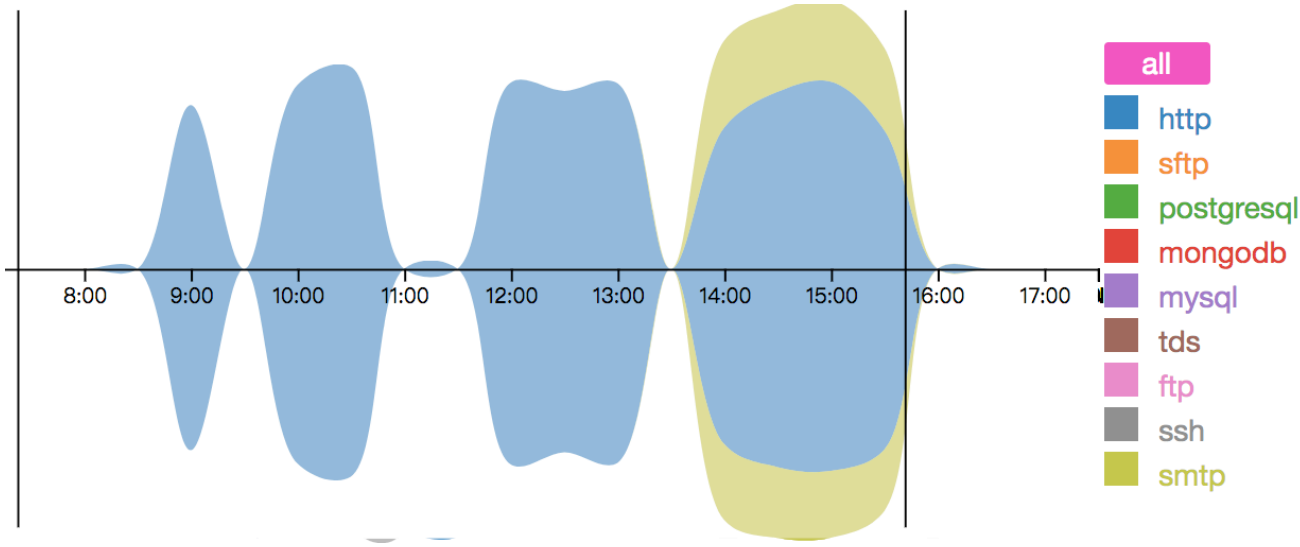

图 1.6

我们通过选取该部门群体，图 1.5 中展现了该部门群体的信息，其中每条线代表该群体中各个员工，线的两端代表上下班时间，我们可以发现该部门上班时间集中于 8 点左右，下班时间集中于 7 点左右，我们再在该群体中随机选择某一员工同时随机选取某一天，通过分析协议

使用流量图可知日常工作中使用电脑主要以网页访问和邮件发送为主，这与想象中也一样，财务部门多以本地操作为主。

人力部门：

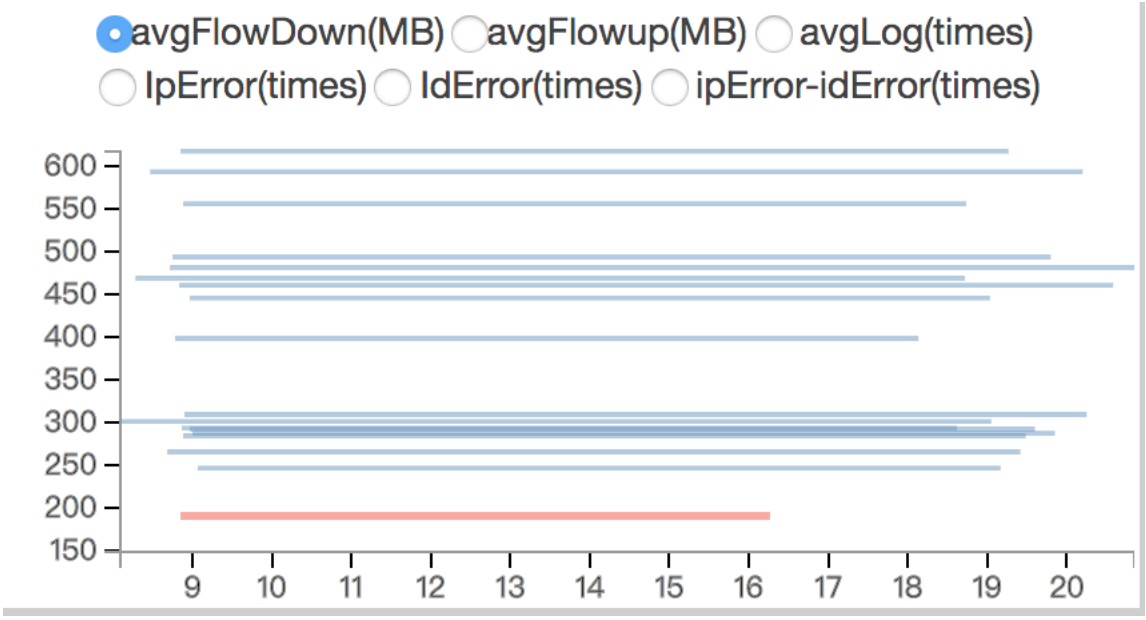

图 1.7

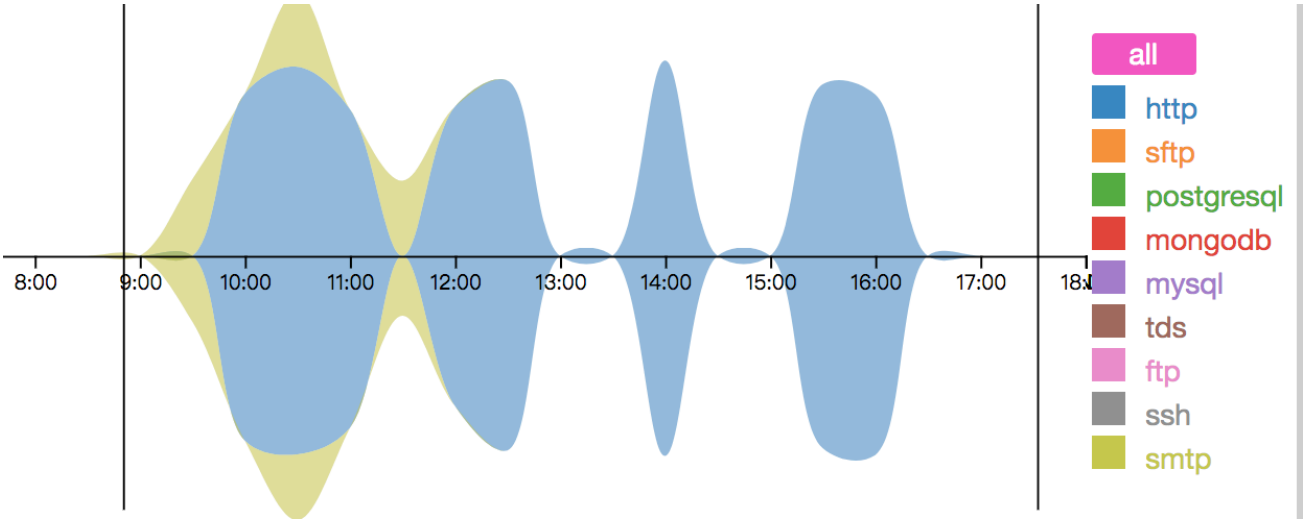

图 1.8

我们通过选取该部门群体，图 1.7 中展现了该部门群体的信息，我们可以发现上班时间多集中于 9 点左右，下班时间多在晚 7 点及以后，其中高亮为红色的是该部门领导，相对于部门领导，下属员工下班时间更晚，上班时间长，随机选取该部门员工以及对比部门领导，通过分析协议使用流量图可发现行为相差无几，使用电脑多为访问网页与发送邮件，这与想象中也差不多，人力部门用电脑的行为比较少。

研发部门：

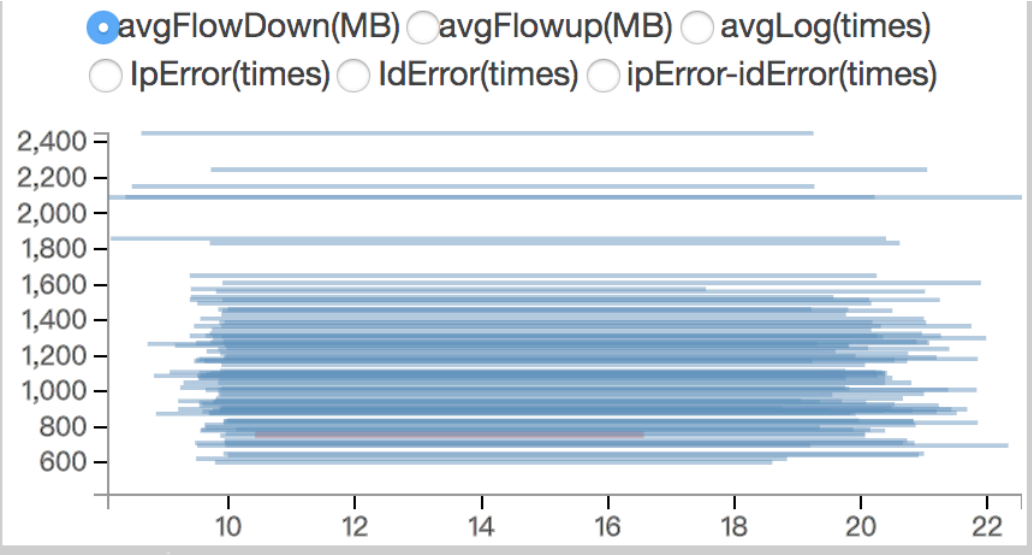

图 1.9

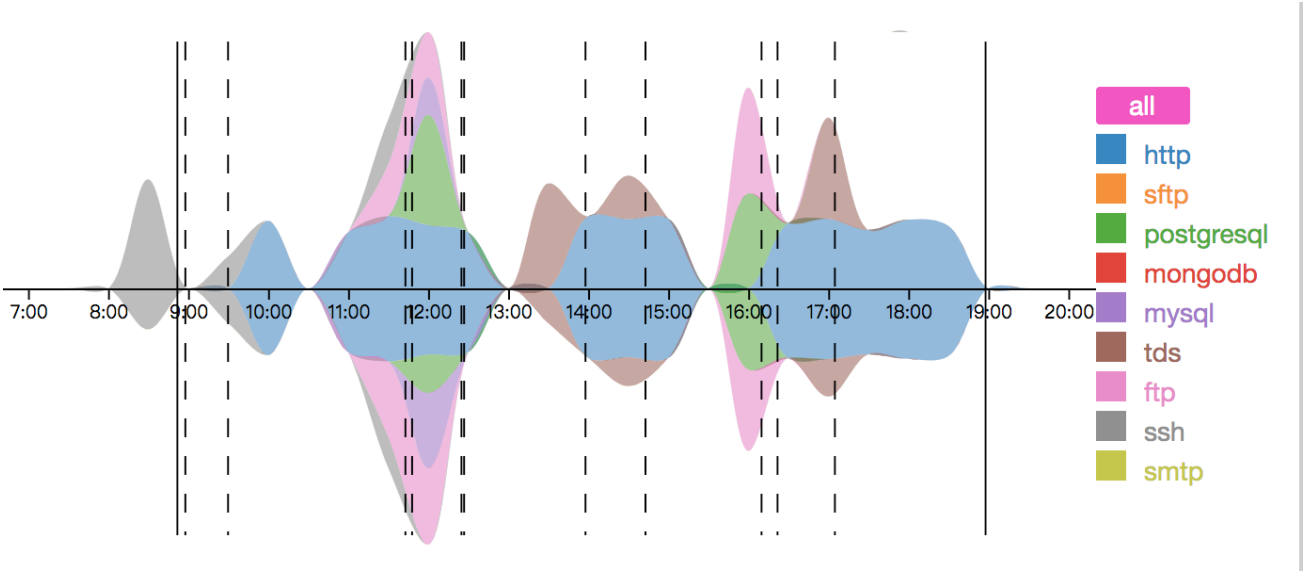

图 1.10

我们通过观察三个研发部门群体，发现日常行为类似，随机选取某一个研发群体，图 1.9 中展现了该研发部门群体的信息，研发部门上班时间多集中于 9-10 点而下班时间多集中于晚 8-10 点，相对于其他部门，上班时间晚下班时间也晚，其中高亮为部门领导，观察发现，相对于部门领导，各个员工上班早下班晚，上班时间明显较长，随机选取部门员工以及对比部门领导，对于协议流量图分析可知，研发部门各个员工与二级领导行为都差不多，对于各种协议都有流量产生，这也在意料之中，研发部门必然会访问数据库服务器、网页访问查资料、邮件沟通需求等等，产生各种协议流量，而部门领导与下属员工的区别在于，员工登陆行为次数更多，部门领导较少（可能只是查看进度遇到问题做决策等等），员工产生的各种协议的流量也会更大，员工作为开发主力这也是必然的。

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. 事件列举

事件 1:

通过对视图 3 的观察我们发现，id 为 1487 号员工的用户的 ip 地址登录错误的的次数明显高于正常范围。（图 3.1）

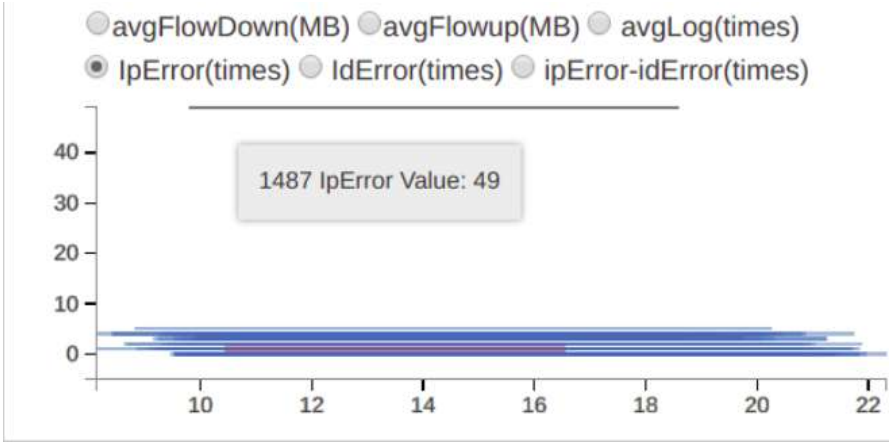

图 3.1 各 ip 地址登录错误次数统计（黑线为 1487）

而当统计 ip 地址登录错误和 id 被登录错误次数的差值时发现，1487 号员工两项统计数据的差值也明显高于正常范围。（图 3.2）

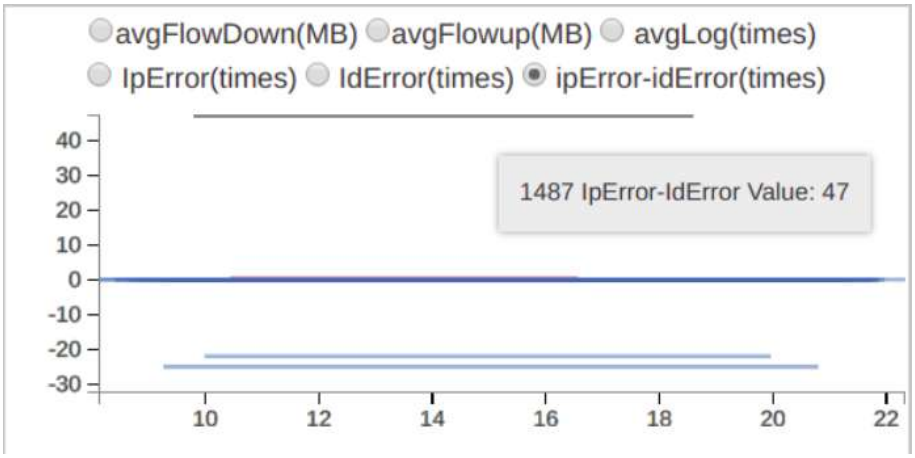

图 3.2 各 ip 地址登录错误次数与各 id 用户名被登录错误次数差值统计

通过分析，我们认为会产生上述情况的原因是：1487 号员工的 ip 地址的错误登录并非用来登录自己的用户 id，而是试图侵入其他员工的用户 id。

事件 2

通过事件 1，我们所定了 1487 号员工为怀疑对像，选中他后，观察视图 4 我们发现该员工在 11 月 4 日在没有产生别的流量协议的情况下产生了 ssh 协议流量。通过联动，我们在视图五中展示了 1487 号员工在 11 月 4 日的各协议流量情况。（图 3.3）

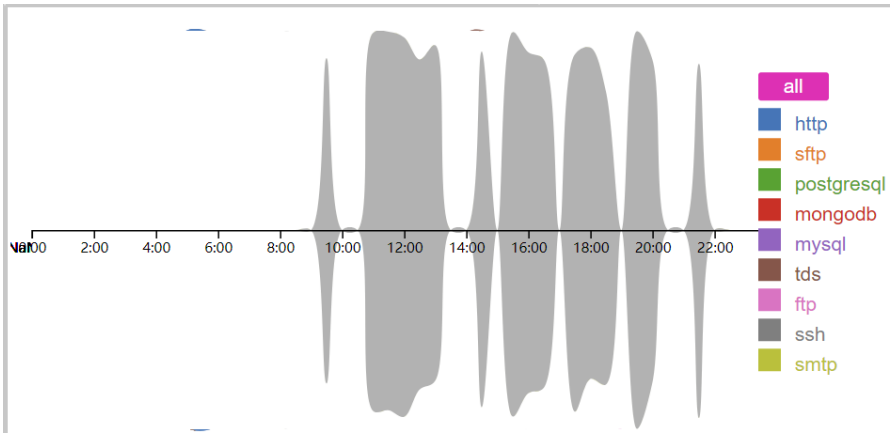

图 3.3 1487 号员工 11 月 4 日流量汇总图

从图 3.3 中我们可以看出，1487 号员工在 11 月 4 日当天没有到公司签到，也没有登录自己的用户 id。并且除了 ssh 流量外没有产生别的协议流量，这基本排除了他远程工作的可能性。因此 1487 号员工极有可能在 11 月 4 日通过自己的 ip 地址使用 ssh 建立连接试图侵入其他员工的用户 id。

事件 3

为了寻找被 1487 号员工在 11 月 4 日侵入的用户 id，我们从视图 3 的统计中，ip 地址登录错误和 id 被登录错误的差值为负数并且明显小于其他员工的员工 id，它们是 1211，1080，1228。（图 3.4）

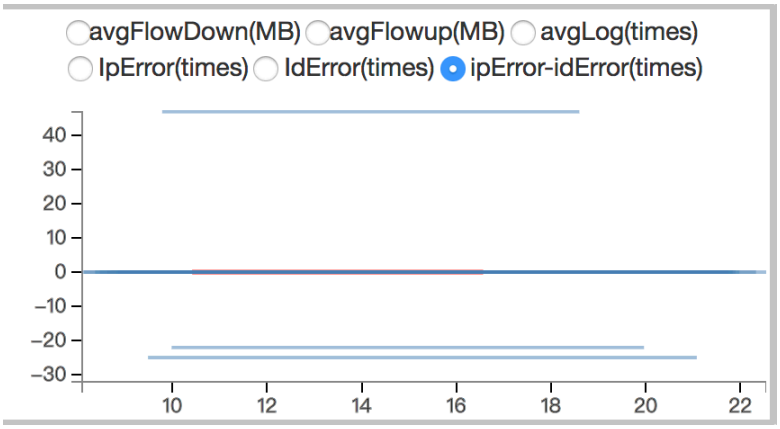

图 3.4 各 ip 地址登录错误次数与各 id 用户名被登录错误次数差值统计

通过观察三位员工在 11 月 4 日当天的流量汇总图，我们发现 1211 号员工的流量在 11 月

4 日存在明显异常。（图 3.5）

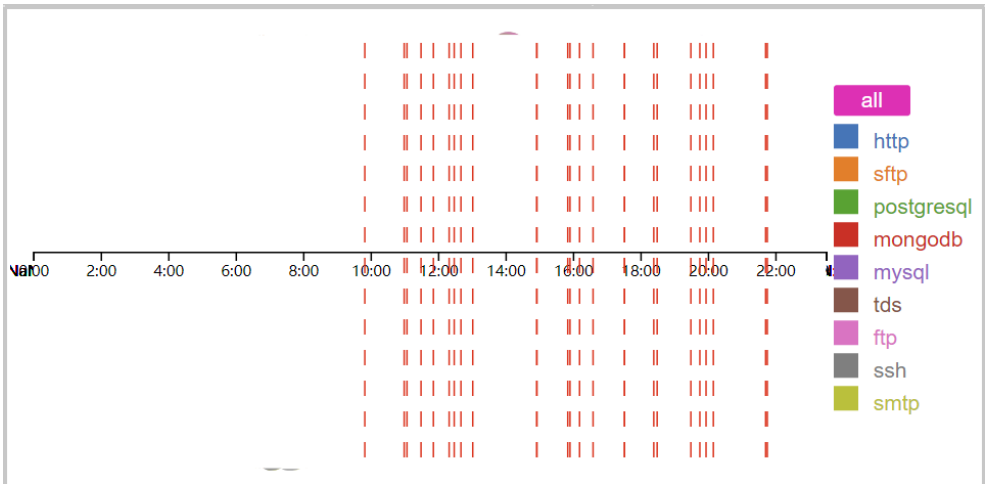

图 3.5　1211 号员工 11 月 4 日流量汇总图

从图 3.5 中我们可以看出，在 11 月 4 日的 10:00 到 22:00，1211 号员工的用户 id 被连续多次试图登录并失败，但 1211 号员工在当天未来到公司也没有产生与工作相关的协议流量。

因此我们猜测在 11 月 4 日，1487 号员工通过 ssh 建立链接，从 10:00 到 22:00 之间一直试图侵入 1211 号员工账号。

事件 4

而对于和 1211 号员工一样账号明显在 11 月中某些时间被人不断试图入侵的员工 1080。我们采用视图 4 和视图 6 联动的方法动态观察他 11 月每一天的流量汇总情况。发现在 11 月 3 日，该员工的账号被多次入侵失败。（图 3.6）

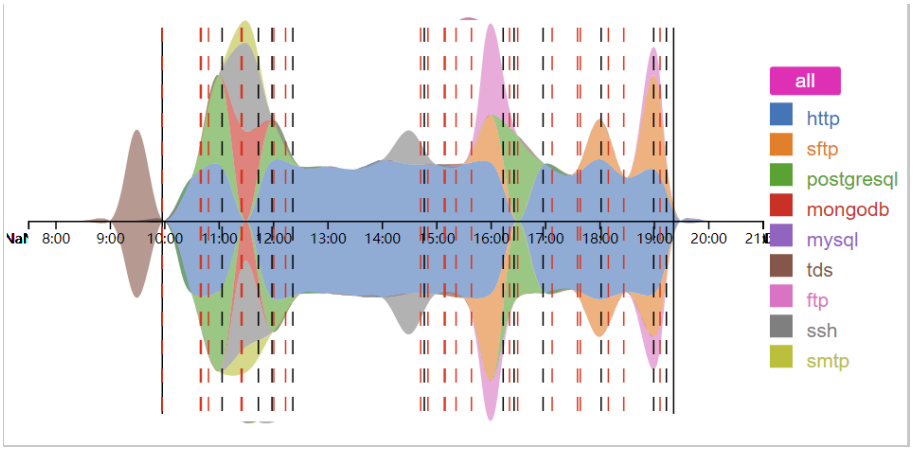

图 3.6　1080 号员工 11 月 3 日流量汇总图

事件 5

同样的对于 1228 号员工，我们在动态观察后发现，他在 11 月 6 日，也出现了账号被多次

入侵失败的异常情况。（图 3.7）

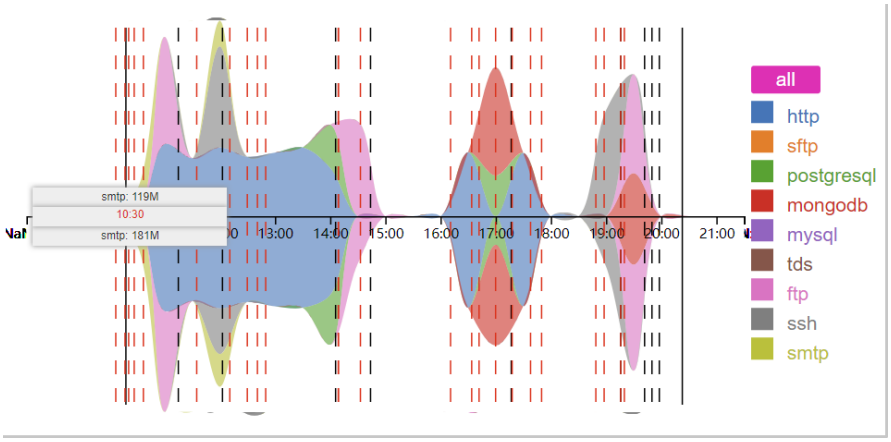

图 3.7　1228 号员工 11 月 6 流量汇总图

事件 6

联动观察后发现，ip 为 1281 的研发部门员工在他的同组成员中，所产生的下行流量明显要多并且落差很大（图中红线为 1281 号员工，蓝线表示同组其他员工）（图 3.8）

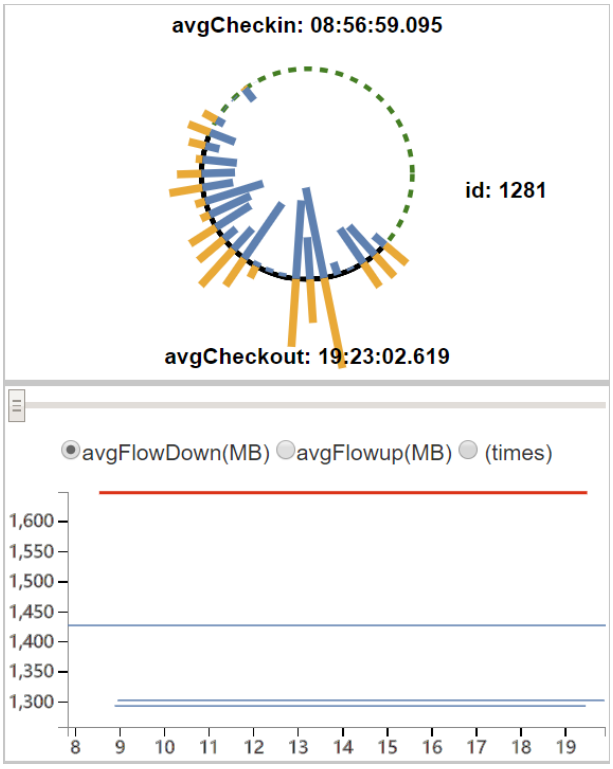

图 3.8　1281 号宇员工平均下行流量图

二．事件关联分析

人物分析:

1、事件 1，2 中试图侵入其他员工账号的 1487 号员工，事件 3，4，5 中被侵入账号的 1211，

1080，1228 号员工，和事件 6 中下行总流量异常的 1281 号员工都属于研发部门。

2、1080，而被侵入账号的 1211，1228 号员工都是隶属 id 为 1059 的研发部门领导下的二级领导。

3、1487 号员工是以 1228 号员工为二级领导的研发部门小组的成员。

数据库定向查询添加的信息

为了得到更多的信息，我们把有问题的员工 id 在后台 mysql 数据库中进行了定向查询。得到了以下信息。

1、试图侵入 1211、1080，1228 号员工账号的 ip 地址为 10.64.105.4，即 1487 号员工的 id。

2、1487 号员工曾经成功访问过他的领导 1228 号员工的账号三次，时间分别为：11 月 6 日 19:42:57；11 月 16 日 20:22:04；11 月 24 日 12:43:41.

在观察 1228 号员工 16 日和 24 日的流量情况后（11 月 6 日的流量汇总图在图 3.7 中已经显示）我们发现在被 1487 号员工入侵的时间里，1228 号员工 id 产生的流量基本为 mysql、mongodb 这样的数据库协议和 ftp 即传递文件协议。（图 3.9，图 3.10）

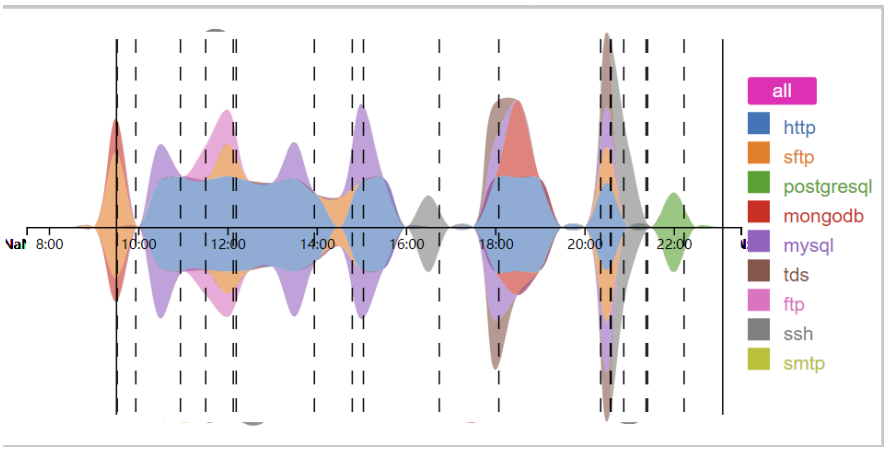

图 3.9 1228 号员工 11 月 16 日流量汇总图

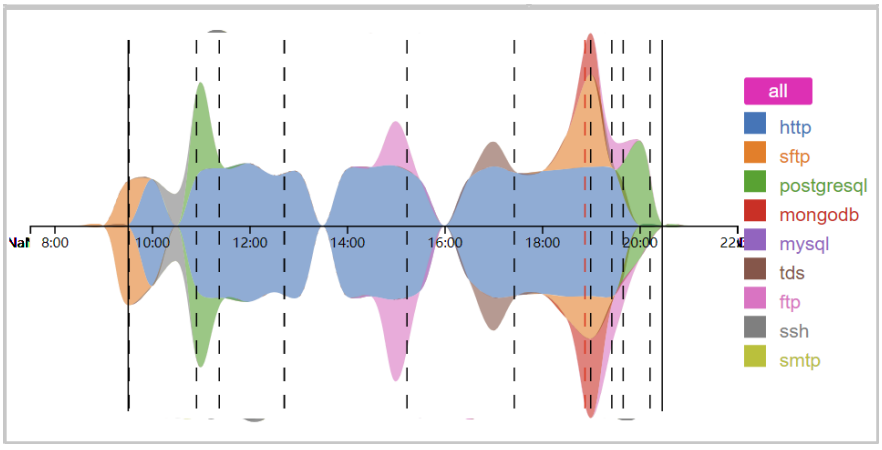

图 3.10 1228 号员工 11 月 24 日流量汇总图

3、 而事件 5 中流量异常的 1281 号员工在我们从数据库中定向查询后发现，他和 1487 号员工同时在 11 月 27 日提出了辞职申请。同一天提出辞职申请。而同一天还有 1376 号员工也提出了辞职，该员工则与同样被是图侵入用户的 1211 号员工在同一小组。

4、1487 号员工的网页浏览协议数据中我们还发闲了其曾经浏览过招聘广告

分析和猜想

被 1487 试图入侵账号的三个员工都是隶属 id 为 1059 的研发部门领导的二级领导。他们的账号中很有可能有关于新产品的信息内容

因此我们揣测，很有可能是 1487 号，1281 号，1376 号员工共同试图窃取公司新产品的信息向外界传递并且在事后后辞职。