2016 年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 2016 - mini challenge 1）

答 卷

参赛队名称： 上海交通大学-樊昕

团队成员： 樊昕，上海交通大学，527309993@qq.com，队长

冯柱天，上海交通大学，zt_feng@sjtu.edu.cn

姜伟鑫，上海交通大学，littlesanjin@sjtu.edu.cn

指导老师： 董笑菊，上海交通大学，xjdong@sjtu.edu.cn

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，python，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： 找出 BigBusiness 公司内部网络中的客户端与服务器，并给出 BigBusiness 公司的网络体系结构拓扑图；对 BigBusiness 内部网络中的服务器进行分类，分类标准不限，比如:按照节点类型、按时间特点、按行为特点、按流量特点等等。（请将回答尽量限制在 1500 个字和 10 张图片内）

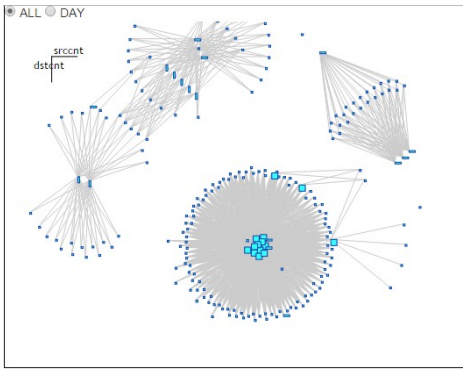
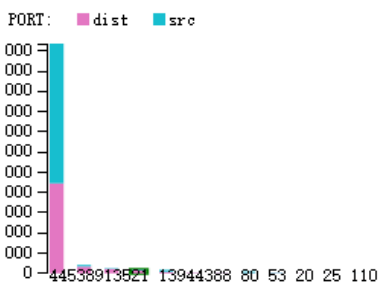


图 1.1

如图 1.1 所示，为整体的网络拓扑图，可以看出，总的网络可以分成 1 个较大的子网络和 3 个较小的子网络。其中集中在各个子网内部的节点 ip 为公司网络内部的服务器。

Ip 10.18.112.246 很重要，因为它发出的通讯量远远大于其他 ip，且大多是通过 445 端口，推测其为公司的主服务器。



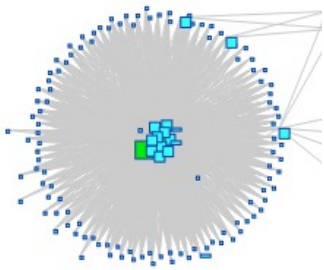


图 1.2　　图 1.3

如图 1.2 所示为典型端口通信总次数的条形图，如果选取某个特定端口，以端口 21 为例，可以看到在图 1.3 中间绿色高亮的 ip：10.18.112.26 通过 21 端口提供了 ftp 服务。

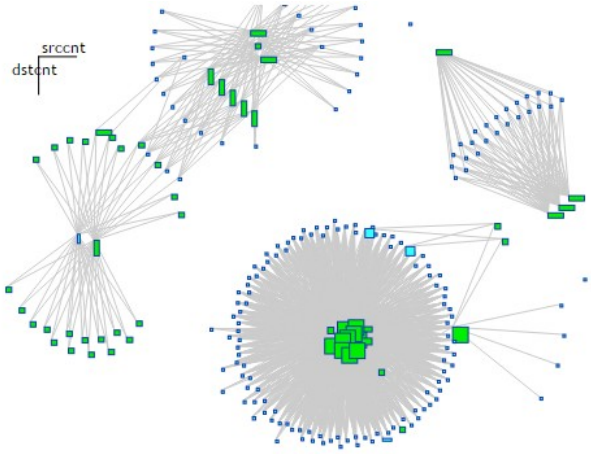


图 1.4

图 1.4 中绿色高亮的点是经过端口 445 的 ip：10.118.216.221，10.18.112.247，10.145.216.221，10.145.216.221，10.46.236.221，10.67.220.221……，我们认为这些 ip 是提供了文件共享服务的服务器。

同理，可以得到 ip：10.145.216.221 通过 389 端口提供轻型目录访问服务。

以此类推，在图 1.2 的视图里选中不同的端口，可以按照提供的服务类型将服务器分类。

挑战 1.2： 找出可能存在的异常通信模式（异常事件），异常标准不限，比如：网络中的周期行为模式和活跃时段变化、某个子网访问或被访问量变化、子网连接模式变化、连接可能传输的文件类型（exe、bmp、jpg 等）的变化等等，建议给出至少 5 种异常通信模式。

（请将回答尽量限制在 1500 个字和 10 张图片内）

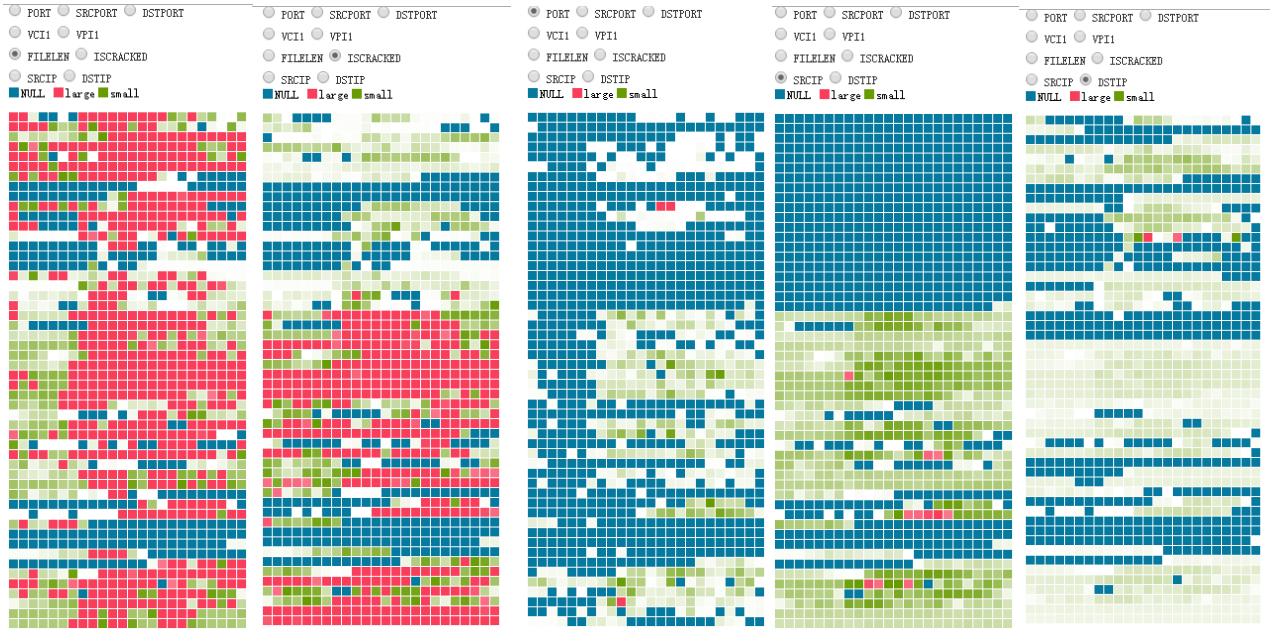


图 2.1　　图 2.2　　图 2.3　　图 2.4　　图 2.5

如图 2.1 所示的矩阵每个方块表示一个小时的数据量，横坐标为 24 小时，纵坐标为 53 天。其中蓝色表示 0，红色表示相对比较大的数据量，绿色处于中间，根据数据量的大小从白色到绿色渐变。

现在我们选中数据类型为 FILELEN，可以得到的结论是：公司的通信一般从早上 7 点到 24 点，符合一般人的作息规律。此外，有一个异常是 9.2 全天和 9.3 前 22 小时没有任何通信数据，可能这两天公司有设备维修或者断电情况。

类似的，选中数据类型为 iscracked 得到图 2.2。可以看出从 8.12 开始有将近一周的时间发出的文件有大量的损坏。

图 2.3 的数据类型为 port，同时在图 1.2 视图中选中端口 21。可以看出 21 端口的活跃时间和 filelen 的时间基本重合。此外，如果查看 21 端口作为目的端口的特性，可以发现 21 端口作为目的端口的次数集中在 7.31 下午 1 点-3 点，点击这个小格，然后在平行坐标中选中 vci 为 selected，如图 2.6。

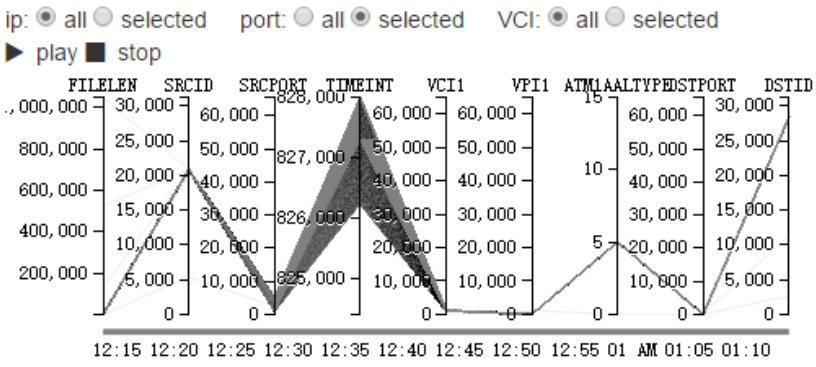


图 2.6

可以发现在这两个小时里从 20834（10.65.220.174）向 28359（10.18.112.26）发出了将近 3 万次的连接，相当异常。

端口 25 有和端口 21 类似的特性。同样，端口 389、135、443、88、80、53 具有特性：活跃时间集中在前几天。端口 20、110 的使用比较分散。端口 139、445 具有类似特性。

在图 1.1 所示视图中选中 ip：10.18.112.246，然后在矩阵视图中将数据类型切换到 srcip，得到图 2.4，为这个 ip 作为来源 ip 的次数分布。切换到 dstip，得到图 2.6，为它作为目的 ip 的次数分布。可以看出这个 ip 在前 20 天左右大都是作为目的 ip，而之后大都作为来源 ip 发出信息，并且在 8.3 中午的时候有个高峰。如果我们想要详细了解这一个小时里的数据情况，可以单击这个小格，观察平行坐标视图，如图 2.7。

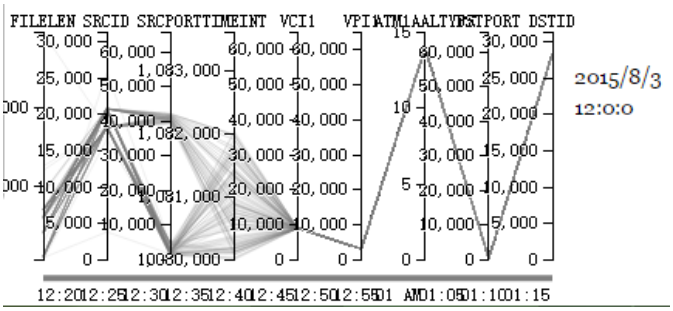


图 2.7

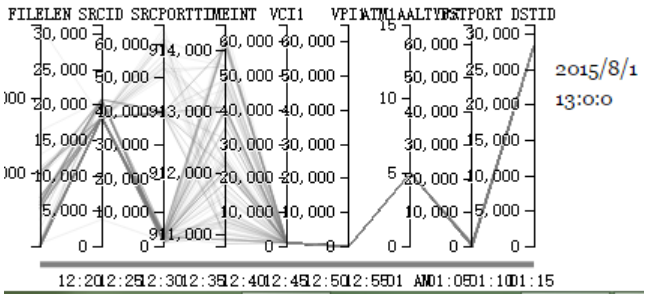


图 2.8

对比其他时间段（图 2.8），将鼠标移动到 path 上查看具体数值，可以看出这个时间段的异常情况为：同样为 dstip 是 10.18.112.246，数据经过的 VCI 不是往常的 1174 而是 9075，主要数据分布在前半个小时，来源端口和 ip 相对集中，虚拟管道的类型也和平时不同，变为 14。

此外，观察不同日期的力导向图也可以查找拓补结构的异常变化。

挑战 1.3： 找出经同一虚拟管道（VPI、VCI）进行通信的源 IP 和目的 IP 的分布规律和连接模式；找出经同一虚拟管道（VPI、VCI）承载的应用分布和变化规律；说明每个（VPI、VCI）虚拟管道的传输数据量在不同时段的变化情况；建议至少给出 5 种虚拟管道通信模式。（请将回答尽量限制在 1500 个字和 10 张图片内）

（下面是挑战 1.3 的答题区域）

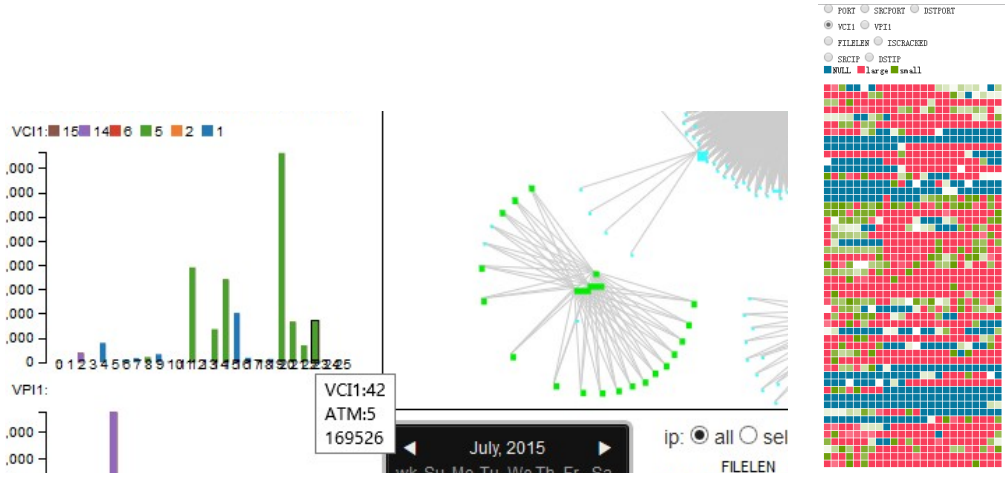


图 3.1　　图 3.2

如图 3.1 所示，选中 VCI1 为 42 的条形图，力导向图视图里会过滤出通过这一管道的 ip，用绿色高亮。说明经过这一管道的 ip 分布为如图所示的子网。主要由中间的服务器为中心发散分布。在矩阵视图里选择 VCI1，可以看出除了蓝色区域的某几天以外都有比较大的数据量通过这一管道。

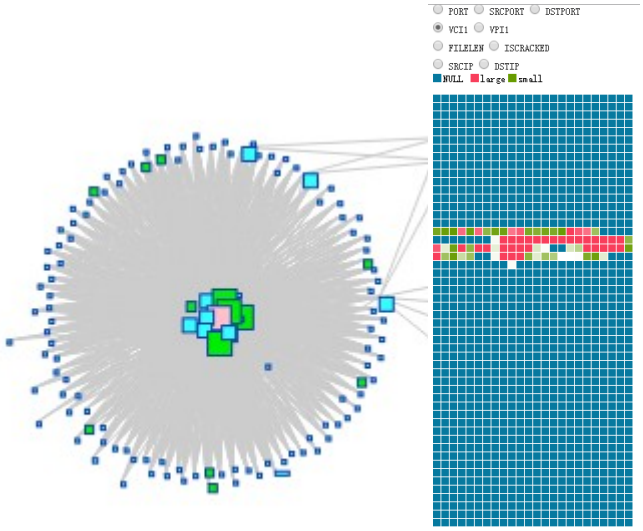


图 3.3

图 3.3 为选中了 VCI1 为 49 的管道，分布如图所示，数据传输量集中在 8.8 开始的四天。

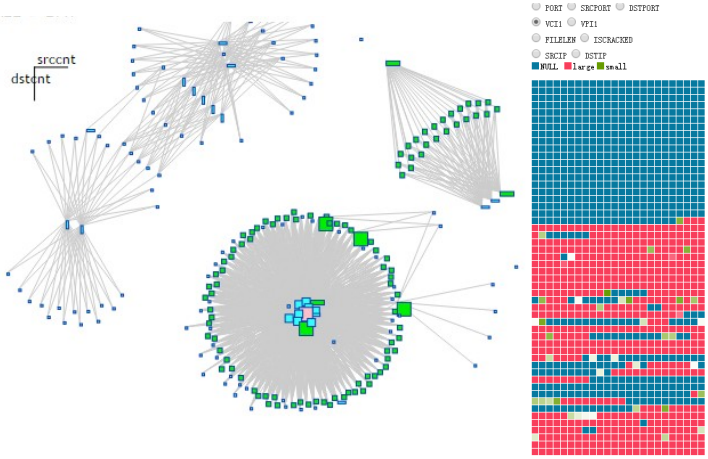


图 3.4

图 3.4 为选中了 VCI1 为 94 的管道，除了大子网的部分通过这个虚拟管道之外还有一个子网也通过这个管道。这个管道的数据量主要分布在后两个月。

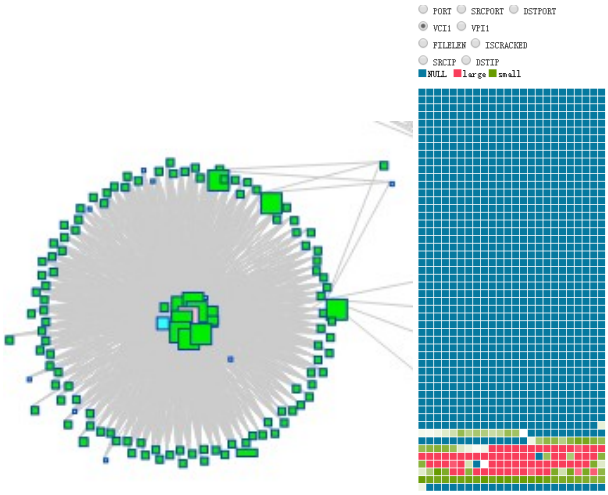


图 3.5

图 3.5 为选中了 VCI1 为 543 的管道，主要用于大子网，数据量主要分布在后最后一周。

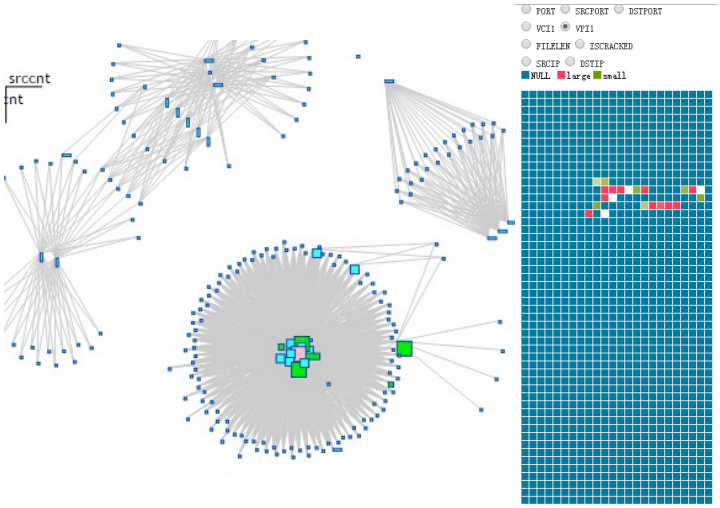


图 3.6

图 3.6 为选中了 VPI1 为 3181 的管道，主要用于大子网向小子网直接通信，数据量主要分布在 8.3 之后的几天。

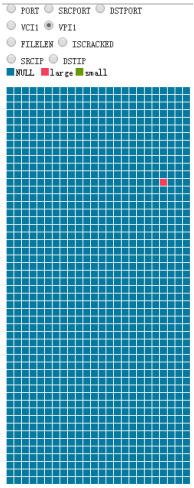


图 3.7

图 3.7 为选中了 VPI1 为 3212 的管道，数据量主要分布在 8.3 晚上 8 点到 9 点。

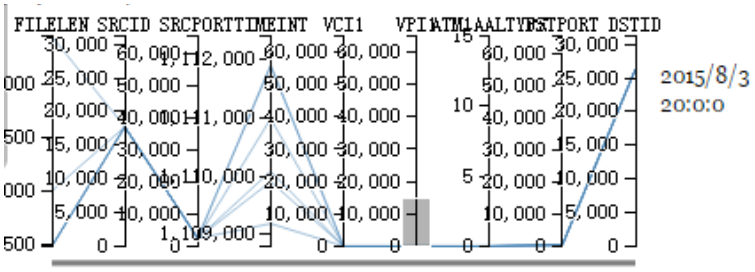


图 3.8

具体数据如图 3.8 所示。

挑战 1.4： WeSuCom 公司要求参赛队设计的可视分析方案能展示数据从源到目的经历应用层、网络层和链路层的全过程，实现对数据连接的多层次可视分析，探索其可能反应的用户行为模式。请参赛队在本题中具体说明其方案是如何满足该需求的。（请将回答尽量限制在 800 个字和 5 张图片内）

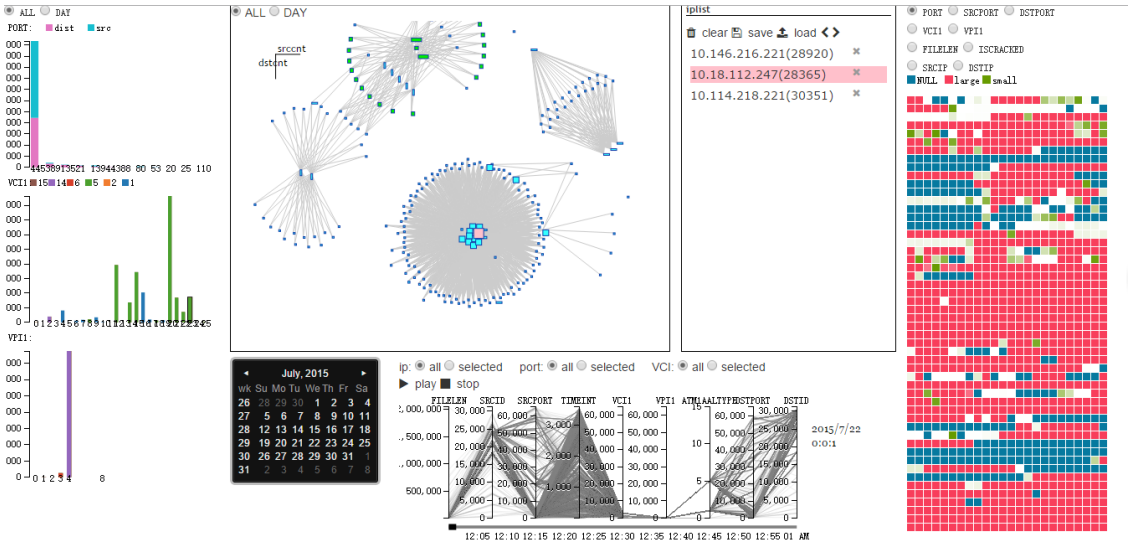


图 4.1

应用层（FILELEN 等）的数据主要体现在矩阵视图。网络层的数据 port 可以通过左侧条形图进行筛选，ip 用力导向图展示。链路层数据主要用于左侧条形图中，针对 ip 的力导视图进行筛选，此外也可以在矩阵视图中显示特定虚拟管道的数据随时间的分布。

主要使用过程为，根据需求，在左侧直方图中选择某一数据，然后在力导向图视图筛选出的高亮 ip 中选择需要的加入到右侧的菜单中，然后在矩阵中观察这个 ip 的收发数据随时间的变化，找到值得观察的某一个小时单击选中，在下方的平行坐标视图中用动画的形式具体查看这一个小时的数据收发情况。通过平行坐标可以看到这段时间内数据从网络层到链路层再到应用层的全过程，可以用于具体分析。