VizBoard - 公司内部威胁情报可视分析

朱肇国, 徐明皓, 植禹衡, 李森, 孙建凯

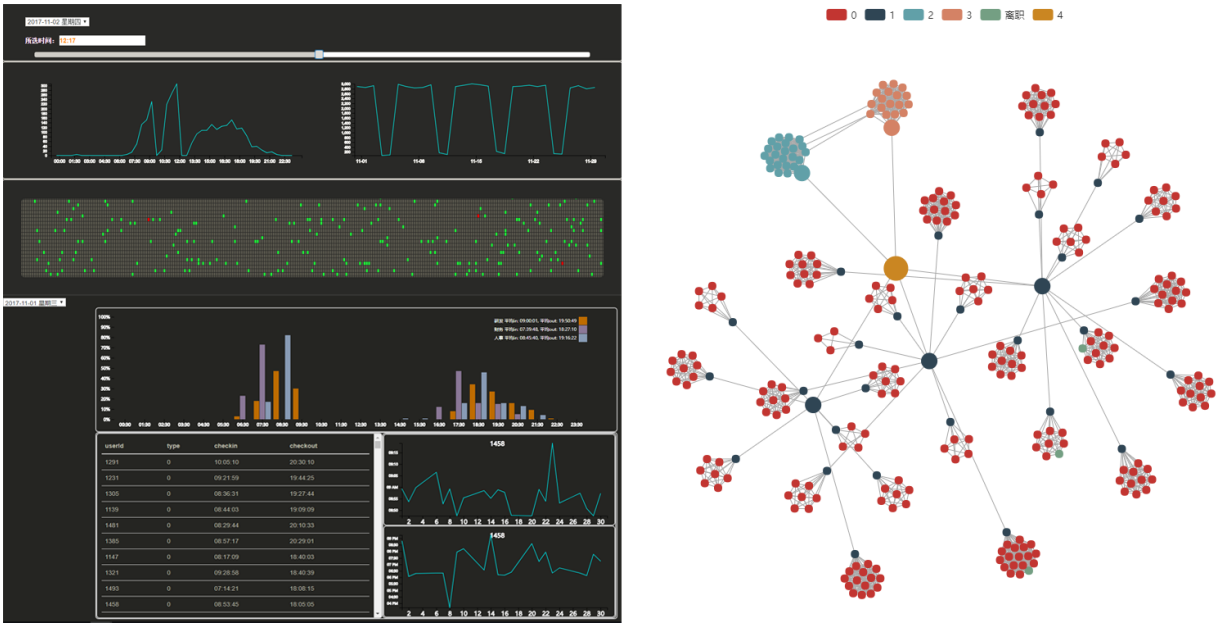


图 1: 按逆时针方向从左上图分别表示某时间段内服务器登录情况, 签到情况, 邮件往来情况

摘要—通过数据理解企业运行是非常重要的方式，可视化和可视分析是理解数据的最佳手段，发觉数据的真正价值在于分析。通过数据进行企业决策，安全形势评估，其核心技术都是只是提取。可视化要求数据本身是完整的、正确的、静态的、清晰的和结构化的，然而复杂数据自身的一些特点也使得它难以被自动分析，包括空间异质性、空间自相关和空间多尺度。我们的项目强调以用户为中心的可视化与可视分析方法，借助 ECharts.js 等可视化库，采用平行坐标图、弦图、像素图、力引导图等方式，通过登录日志、网页访问日志、TCPLOG 日志、邮件日志、打卡日志，分析了企业的财务部门、人力资源部门、研发部门的组织结构、日常工作行为、异常事件及其关联。

关键字—大数据, 安全, 可视化, 数据分析, 可视分析

1 介绍

我们的论文展示了 ChinaVis 2018 的结果。我们的目的是分析一家互联网高科技公司 HighTech，有财务、人力资源和研发三个部门。我们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。我们的项目通过可视分析技术将计算智能与人类智慧紧密结合，设计并实现了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。采用平行坐标图、弦图、像素图、力引导图等方式，分析企业各部门的组织结构、日常工作行为、异常事件及关联。

- 朱肇国. 上海交通大学. E-mail: armando@sjtu.edu.cn.
- 徐明皓. 上海交通大学. E-mail: xuminghao118@sjtu.edu.cn.
- 植禹衡. 上海交通大学. E-mail: zyh1996@sjtu.edu.cn.
- 李森. 上海交通大学. E-mail: 13162054619@163.com.
- 孙建凯. 上海交通大学. E-mail: jiankai@sjtu.edu.cn.

2 我们的方案及其特点

2.1 平行坐标图

平行坐标对多维数据的表达实现了多维数据在二维平面上的表示。其思想就是将 N 维数据点映射到处于 N 条平行的坐标轴上的彼此相连的 $N-1$ 条线段。这 $N-1$ 条线段与 N 条轴相交的 N 个点分别代表了数据点的 N 维数据。这条代表 N 维数据的折线可用 $N-1$ 个线性无关的方程所表示 1:

$$\frac{x_1-a_1}{u_1}=\frac{x_2-a_2}{u_2}=\cdots=\frac{x_n-a_n}{u_n} \tag{1}$$

$$x_{i+1}=m_i x_i+b_i,\quad i=1,2,\cdots,n-1 \tag{2}$$

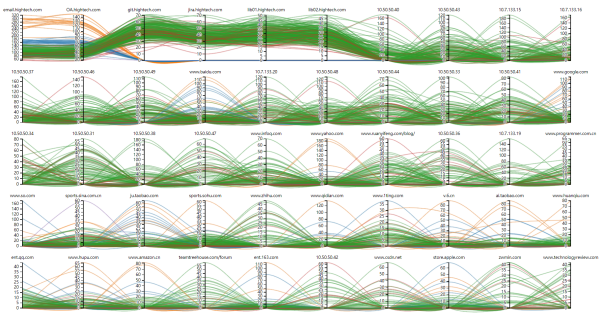


图 2: 平行坐标图分析

其中，$m_i = u_{i+1}/u_i$ 表示斜率，$b_i = (a_{i+1} - m_i a_i)$ 表示在 $x_i x_{i+1}$ 平面中 x_{i+1} 轴上的截距。

2.2 弦图

弦图 (Chord Diagram) 可以显示不同实体之间的相互关系和彼此共享的一些共通之处，非常适合用来比较数据集或不同数据组之间的相似性。节点围绕着圆周分布，点与点之间以弧线或贝塞尔曲线彼此连接以显示其中关系，然后再给每个连接分配数值（通过每个圆弧的大小比例表示），也可以用颜色将数据分成不同类别，有助于进行比较和区分，线的粗细表示权重。我们用弦图表示一个时间段内两个 IP 地址之间的流量。

2.3 力导引图

力引导布局 [1] 可以减少布局中边的交叉，尽量保持边的长度一致。Fruchterman-Reingold 算法，加入点之间的静电力，通过计算系统的总能量并使得能量最小化，从而达到布局的目的。对于图中，节点 i 和 j，用 $d(i,j)$ 表示两个点的欧式距离，$s(i,j)$ 表示弹簧的自然长度，k 是弹力系数，r 表示两个点之间的静电力常数，w 是两个点之间的权重。弹簧模型如公式 3,

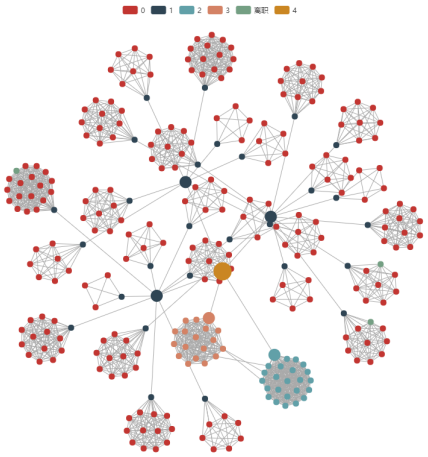
$$E_s = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}k(d(i,j) - s(i,j))^2, \tag{3}$$

能量模型如公式 4,

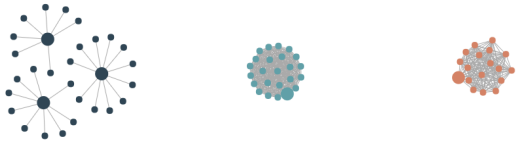
$$E = E_s + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{rw_i w_j}{d(i,j)^2}, \tag{4}$$

3 实验及结论

从邮件往来角度分析，整个的组织架构见图 3a，中间黄色的大圆代表公司最高领导，他直接管理中层领导。中间大小的蓝色圆是研发部部长，小一点的蓝色圆是小组长，红色的小圆是组员，绿色的小圆是离职的员工，之间的连线表示邮件往来。



(a) 所有邮件数据的可视化



(b) 研发部中层　(c) 财务部职员　(d) 人资部职员

图 3: 邮件数据分类别可视化结果

3.1 财务部门

财务部人员组成如图 3c所示，可见财务部有一位主管，所有普通员工都向该主管汇报。

3.2 人力资源部门

人力资源部人员组成如图 3d所示，可见人力资源部中也有一位主管。从图 2得知，人力资源部门对门户网站（如 yahoo，sohu，huanqiu，hupu，163），购物网站（如 taobao，amazon）的访问频率远高于其他部门。

3.3 研发部门

从图 2可知, 公司代码相关的 git 服务器和库服务器等，全是是研发部在访问。中层管理人员分为两级（如图 3b所示），有小组长和研发部部长，普通员工以小组形式存在。

研发部门存在的异常现象比如 5 位员工 1265 1325 1175 1173 1444 通常在非常晚的时间登录服务器，单日下载量大，而且对某个内部服务器网页记录访问量大。

参考文献

[1] P. Eades. A heuristic for graph drawing. Congressus numerantium, 42:149–160, 1984.