公司内部威胁情报分析

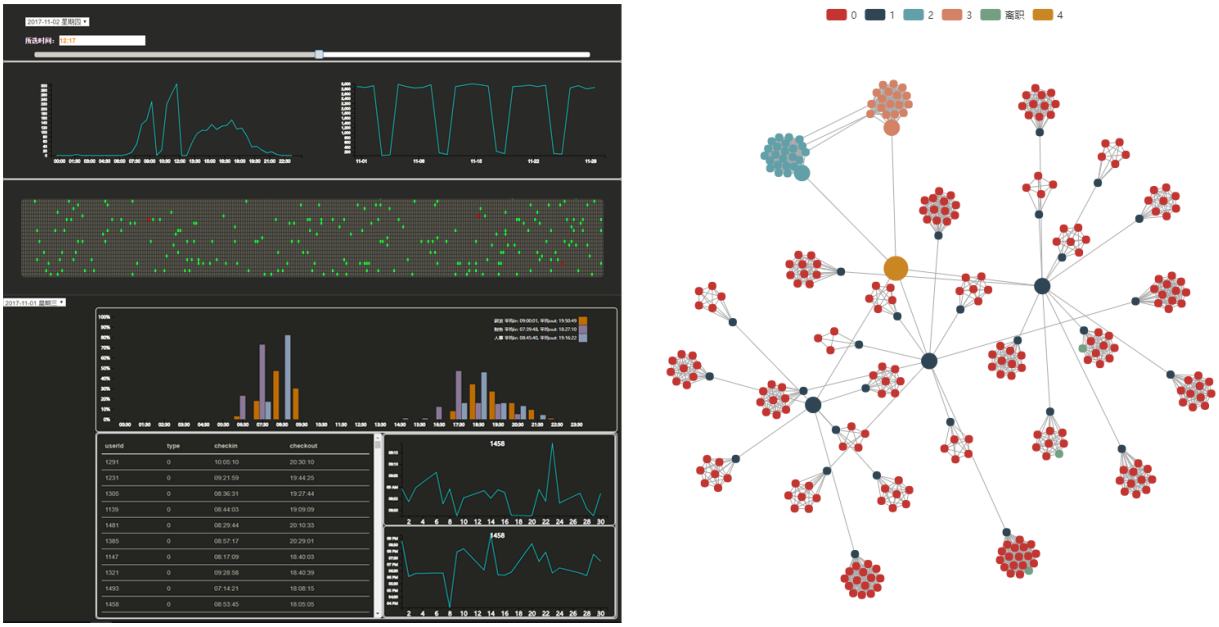


图 1: 概要：按逆时针方向从左上图分别表示 xx, 签到情况, 邮件往来情况

Abstract—

Index Terms—大数据, 安全, 可视化, 数据分析, 可视分析

1 介绍

我们的论文展示了数据可视分析挑战赛的结果。我们的目的是分析一家互联网高科技公司 HighTech，有几百名员工，分属财务、人力资源和研发三个部门。公司高层决定临时成立内部威胁情报分析小组，该小组将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。在分析威胁情报过程中，数据的复杂性需要计算智能处理，但发现和处置安全威胁需要人的经验、认知和判断，可视分析技术能将计算智能与人类智慧紧密结合，是威胁情报人员高效分析和理解威胁情报数据的利器。假设您是威胁情报分析小组的成员，我们的项目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。

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2 我们的方案

2.1 平行坐标图

平行坐标图将数据表中的每一行映射为线或剖面。某行的各个属性由线上的点表示。平行坐标图中的值始保持正常化，最低值绘制为 0%，最高值绘制为 100%。这表示对于沿 X 轴的每个点来说，相应的列中的最低值沿 Y 轴被设置为 0%，此列中的最高值被设置为 100%。各列的刻度完全独立，因此不要将某一列中曲线的高度与其他列中曲线的高度进行比较。

平行坐标对多维数据的表达是数据可视化的重要方法之一。它实现了多维数据在二维平面上的表示。利用平行坐标对数据进行分析处理的技术已经取得了很大的进展, 如刷 (Brushing) 技术、交换坐标轴、抽象等。这些分析技术已经应用到数据挖掘的很多领域, 尤

其在聚类分析中, 平行坐标对数据集的定性分析使聚类结果的合理性得到证明。

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, 方程公式 1:

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

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

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），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，因此这种图表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当中关系，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）。此外，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。

我们用弦图表示了一个时间段内两个 IP 地址之间的流量。

2.3 力导引图

力引导布局最早由 Peter Eades 在 1984 年的“启发式画图算法”一文中提出，目的是减少布局中边的交叉，尽量保持边的长度一致。此方法借用弹簧模型模拟布局过程：用弹簧模拟两个点之间的关系，受到弹力的作用后，过近的点会被弹开而过远的点被拉近；通过不断的迭代，整个布局达到动态平衡，趋于稳定。其后，“力引导”的概念被提出，演化成力引导布局算法 FR（Fruchterman-Reingold 算法）——丰富两点之间的物理模型，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。这种改进的能量模型，可看成弹簧模型的一般化。

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

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

力引导图的伪代码如算法 1:

Algorithm 1: 力引导布局伪代码

```
设置节点的初始速度为 (0,0)
设置节点的初始位置为任意但不重叠的位置
for 总动能 =0; 对每一个节点 i；进行循环 do
    静力 f = (0,0)
    for 对每一个出该节点外的每个节点 j do
        静力 f= 静力 f+j 节点对应 i 节点的库仑斥力
    for 对该节点上的每个弹簧 s do
        静力 f= 静力 f+ 弹簧对该节点的胡克弹力
    该节点速度 =（该节点速度 + 步长 * 净力）* 阻尼
    该节点位置 = 该节点位置 + 步长 * 该节点速度
    总动能 = 总动能 + 该节点质量 *（该节点速度）2
```

3 实验及结论

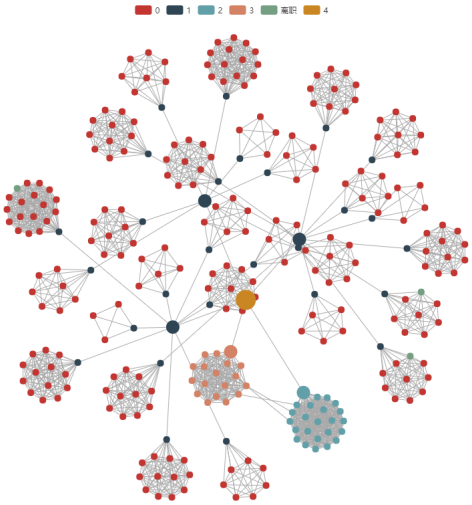
从邮件往来角度分析，整个的组织架构见图 2a，中间黄色的是公司最高领导，只有一位，他直接管理中层领导。中间大小的蓝色圆是研发部部长，小一点的蓝色圆是小组长，红色的小圆是组员，绿色的小圆是离职的员工，之间的连线表示邮件往来。图 2a提供了所有数据的概览。

3.1 财务部门

财务部人员组成如图 2c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 2d所示，可见人力资源部中也有一位主管。从图 3得知，人力资源部门对门户



(a) 所有邮件数据的可视化

(b) 研发部中层管理

(c) 财务部所有职员

(d) 人力资源部所有职员

图 2: 邮件数据分类别可视化结果

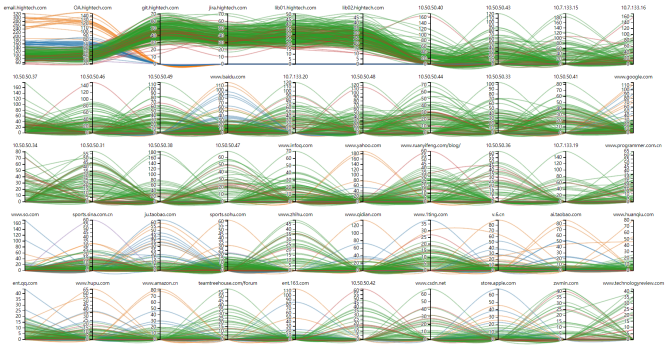


图 3: 平行坐标图分析

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 3, 公司代码相关的 git 服务器、jira（JAVA 管理系统）和库服务器等，全是研发部在访问。中层管理人员的情况如图 2b所示，可见中层管理人员分为两级，有小组长和研发部部长。观察所有组员的信息，可以发现，员工有组成小组的趋势。

3.4 特点