一种园区网风险账号的评估方法

作者：曾煌尧 李丹丹 马严 丛群

北京邮电大学--网络技术研究院

目 录

CONTENTS

背景

研究背景与现状

解决方案

风险账号评估方法

整体思路

方法描述

模拟实验

实验数据

实验结果

01 背景

- ✓ 研究背景与现状
- ✓ 解决方案

研究背景与现状

背景

- 园区网的入网账号通常欠缺安全保护措施。比如初始密码简单、明文传输导致泄露等等原因给园区网账号带来安全风险。
- 用户被盗用之后无法获知

现状

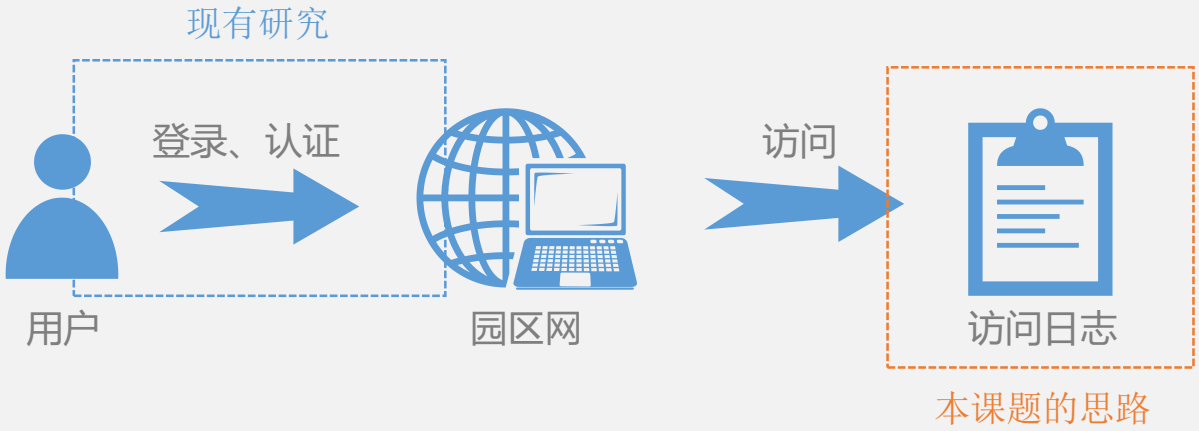
- 现有检测与评估措施：
  1. 双因素验证

     没有搭配双因素认证的园区依然存在风险
  2. 对登录日志进行分析

     登录日志特征维度有限
  3. 对统一身份认证日志进行分析

     特征维度有限、系统普及率有限

解决方案



02 风险账号评估方法

- ✓ 整体思路
- ✓ 详细步骤

整体思路

a. 从**用户设备访问URL的行为**是否有异常的角度出发

b. 基于访问次数序列在**时间**和**空间**上的变化规律量化并提取表现特征

c. 利用**聚类**标记和**余弦相似度，**找出**访问行为异常**的设备和**偏离用户本身访问喜好**的设备（统称风险设备）来进行风险账号的评估

方法描述

特征提取

- **输入**

  用户设备URL访问日志

- **输出**

  用户设备对于URL的访问情况量化的一阶张量 $[D_{std}, A_{risk}, V_{per}, L_{risk}]$
  以及用户设备URL访问相似度 C_{mean}

- **处理流程**

  1. 设备出现次数离散度 D_{std}
     - 风险设备的访问主要集中在某个时间段，并且访问次数会比平均值多
     - 假定$[(t_1, n_1), (t_2, n_2), ..., (t_m, n_m)]$ 为设备在一段时间序列内的访问次数，则

$$D_{std} = std([(t_1, n_1), (t_2, n_2), ..., (t_m, n_m)])$$

方法描述

特征提取

2. 设备多账号风险度 A_{risk}

- 盗号者往往在一个设备上登录过多个账号（至少两个），而正常用户一般在自己的设备上登录的是自己的账号。
- 设 $A_{\max}$ 和 $A_{\min}$ 分别为用户设备在URL访问次数的时间序列 $[(t_1,n_1),(t_2,n_2),\ldots,(t_m,n_m)]$ 里登录过的账号数的最大值和最小值
- 那么用归一化的方式量化这个风险程度即为：

$$A_{risk}=(\sum_{i=0}^{m}\frac{A_{ni}-A_{\min}}{A_{\max}-A_{\min}})/m$$

方法描述

特征提取

3. 收费网络占比V_{per}

- 盗号者的设备往往通过盗取正常用户的账号来盗取收费流量。
- 设设备访问URL的总次数为C_{all}，以收费网络进行访问的次数为C_{paid}
- 则收费网络占比V_{per}为:

$$V_{per} = \frac{C_{paid}}{C_{all}}$$

方法描述

特征提取

4. 对立位置风险度 L_{risk}

- 盗号者的设备往往与正常用户的设备在不同的地点登录。
- 通过分析不同地点的关系，我们可以整理出用户下的设备登录地点的对立关系
- 统计用户账号下出现的对立位置集合 $L_{\mathrm{p}} = \{P_1, P_2, ..., P_n\}$
- 设第 i 次访问URL时处于对立位置的情况为 L_{di}

$$L_{di} = \begin{cases} 1, & d_i \in L_p \\ 0, & d_i \notin L_p \end{cases}$$

- 则对立位置风险度为：

$$L_{risk} = (\{L_{d1}, L_{d2}, ..., L_{dn}\}) / n$$

方法描述

聚类分析

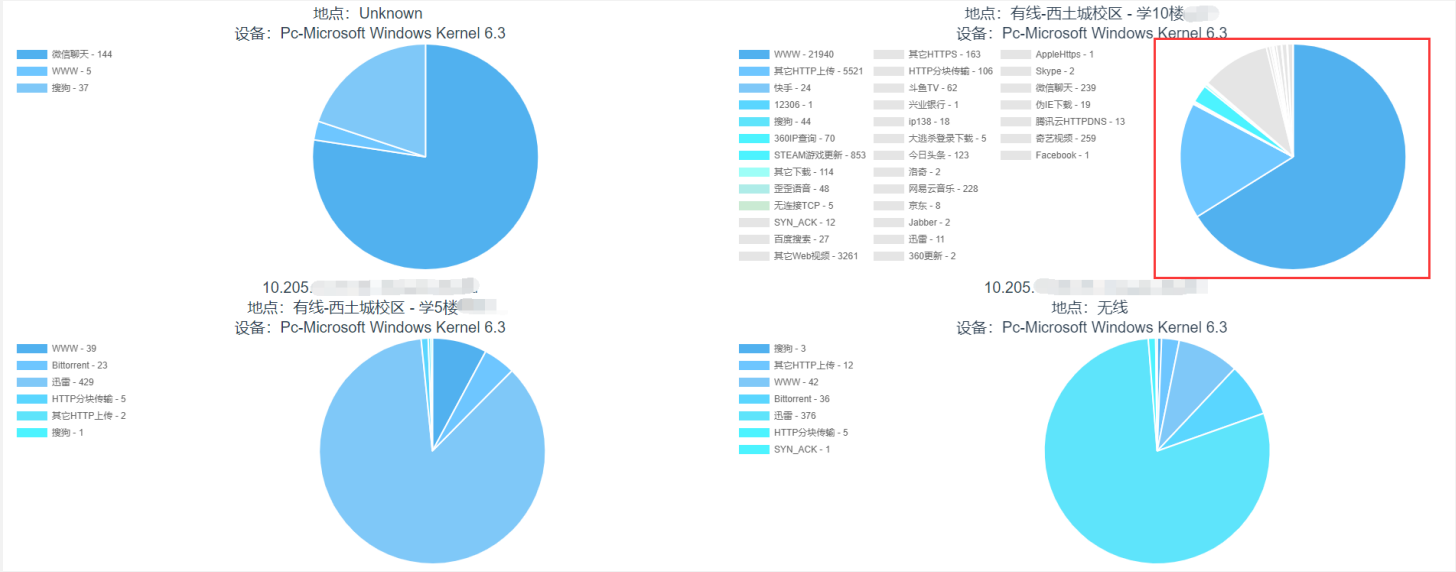
- 将上述特征向量 $[D_{\text{std}}, A_{\text{risk}}, V_{\text{per}}, L_{\text{risk}}]$ 作为聚类分析输入，目的是划分出正常设备的访问和风险设备的访问两个簇

- 由于特征样本正常数量很多，异常样本数量很少，实验次数足够大时二项分布符合高斯分布，故而使用高斯混合聚类

- 得到聚类结果 Y，得出样本是风险设备访问的概率

方法描述

访问相似度计算



风险账号下风险设备的URL访问喜好与正常设备差别过大

方法描述

访问相似度计算

设备访问URL相似度 C_{mean}

- 通过分析用户设备本身访问URL情况，可以发现盗号者的设备和正常用户的设备访问偏好往往会出现较大差异
- 利用推荐算法里的常用的余弦相似度来计算设备间访问URL的相似度能得到较好的量化效果
- 设一个账号下的PC设备列表为 $F=\{F_1,F_2,...,F_n\}$，这些设备访问过的URL对应的标签列表为 $B=\{B_1,B_2,...,B_m\}$。统计每台设备在对应标签下的访问次数，量化成 $n\times m$ 的矩阵。
- 利用余弦相似度计算公式 $$\cos(\theta)=\frac{\sum_{i=1}^{m}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{m}(x_i)^2}\times\sqrt{\sum_{i=1}^{m}(y_i)^2}}$$

方法描述

访问相似度计算

- 经过上述计算之后，可以得到一个$n \times n$ 的矩阵，每一行的元素是：

$$C_i = [\cos_{i,j}, \cos_{i,2}, ..., \cos_{i,n}] \quad (1 \le i \le n)$$

- 由此可以得出每台设备相比于其他设备的URL访问余弦相似度

$$C_{mean} = \frac{-1 + \sum_{i=1}^{n} \cos_{i,j}}{n}$$

当设备访问URL的喜好发生偏离时，对应的 C_{mean} 也会随之降低，风险也会增大。

方法描述

风险评估

- **输入**

  已标注的样本集合S1 + 已计算的用户的设备URL余弦相似度

- **输出**

  风险账号评估阈值

- **处理流程**

  - 对于每个用户，找出其账号下风险设备的最大概率值Y，以及设备列表中余弦相似度最小值Cmean。对于他们进行如下运算：

$$R = \frac{Y + (1 - C_{mean})}{2}$$

当R大于一定阈值的时候，该账号具有风险性。

03 模拟实验

- ✓ 实验数据
- ✓ 实验结果

实验数据

来源：北京邮电大学04.01 —04.10 期间校园网出口的 URL 访问日志

1. 数据源

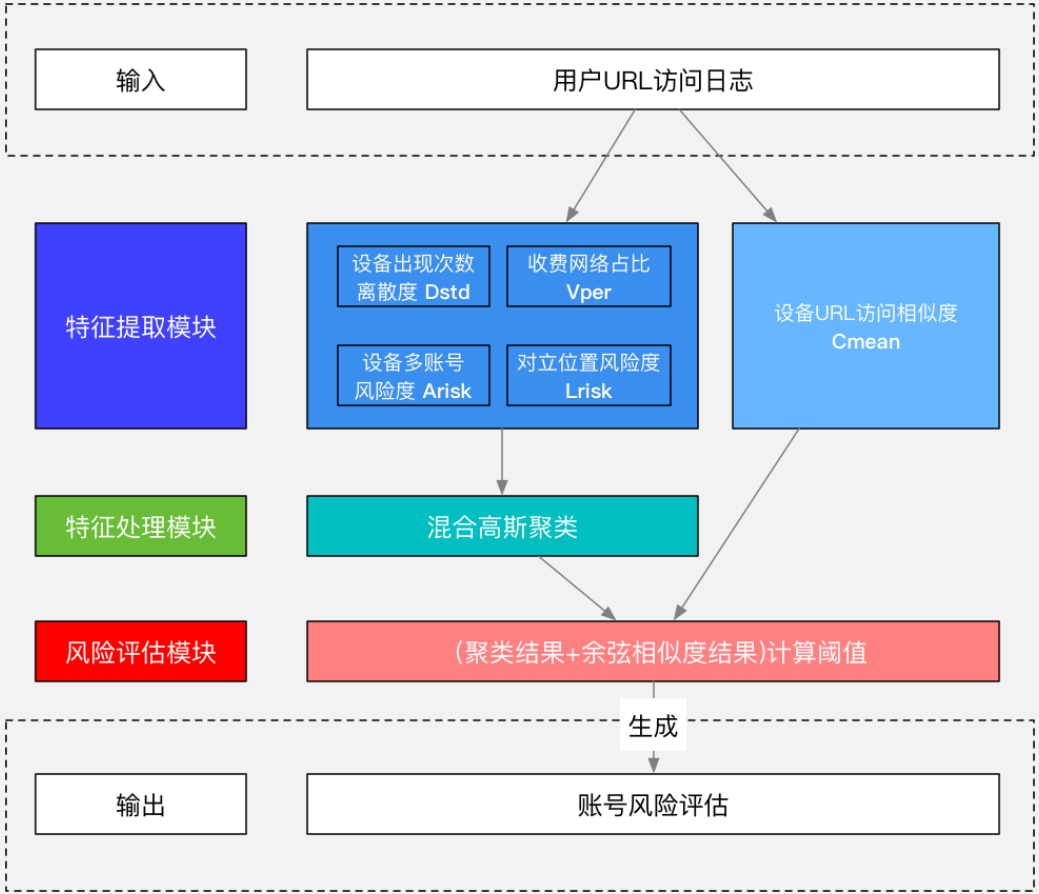
样本类型	样本组成
白样本	980个账号十天内的 URL 日志
黑样本	20个被盗账号十天内的 URL日志（包括10个模拟账号）

表3 用户访问集合S的样本组成

2. 训练与测试集

以7:3的比例将数据分为训练集与测试集，共13.7GB

架构设计


输入
用户URL访问日志
特征提取模块
设备出现次数
离散度 Dstd
收费网络占比
Vper
设备多账号
风险度 Arisk
对立位置风险度
Lrisk
设备URL访问相似度
Cmean
特征处理模块
混合高斯聚类
风险评估模块
(聚类结果+余弦相似度结果)计算阈值
生成
输出
账号风险评估

实验结果（一）

特征表现

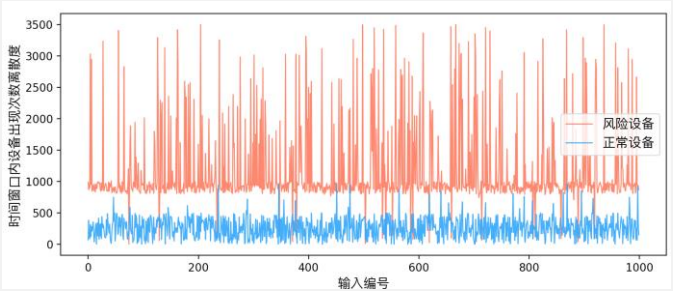


图 1 随机抽取的 2 000 个访问在**设备出现次数离散度**上的取值情况

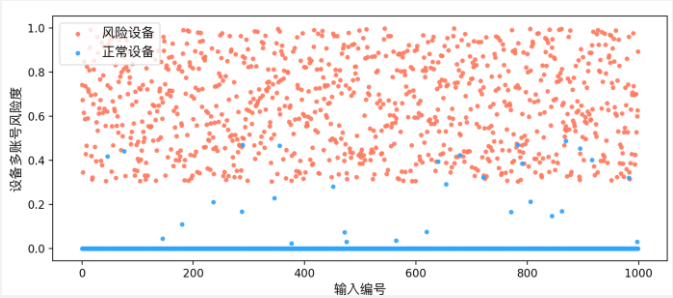


图 2 随机抽取的 2 000 个访问在**设备多账号风险度** 上的取值情况

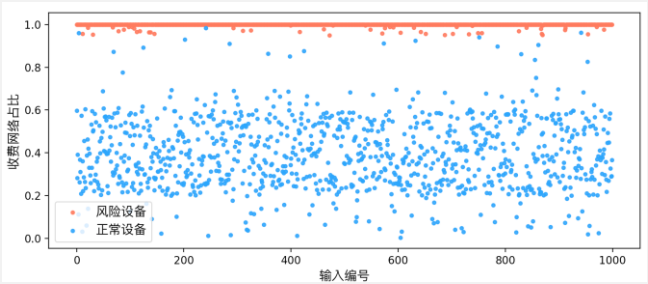


图 3 随机抽取的 2 000 个访问在**收费网络占比**上的取值情况

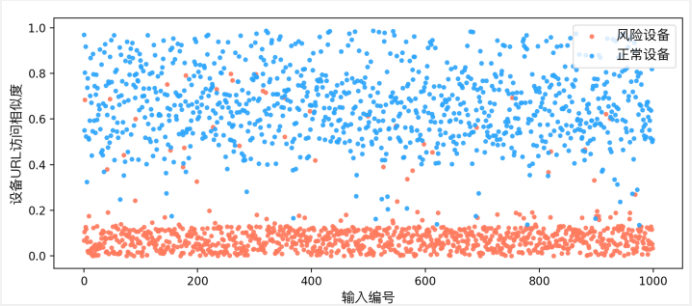


图 4 随机抽取的 2 000 个访问在**设备URL访问相似度**上的取值情况

实验结果（二）

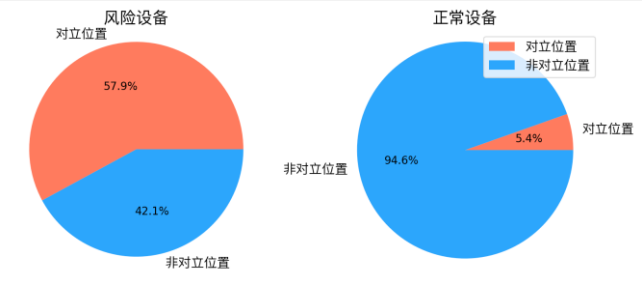


图 5 随机抽取的 2 000 个访问在**设备访问URL对立位置风险度**上的取值情况

左图效果并不十分明显，考虑主要原因是有些风险账号的设备访问IP为非对立位置或者难以判断的对立位置。

识别准确率

共在1000个输入的账号里检测出17个风险账号（总共20个风险账号），准确率为85%

总结

- 不再着眼园区网登录、认证过程，而采用分析账号URL访问日志来评估账号的风险
- 关注于风险设备的发现，量化风险设备相应的特征
- 采用高斯混合聚类与余弦相似度用于发现风险设备从而对账号做出风险评估
- 1000个输入的账号里检测出17个风险账号（共20个）准确率85%

谢　谢　聆　听