北京大学
PEKING UNIVERSITY

基于知识图谱的邮件系统异常用户检测

金建栋
北京大学计算中心
2019.11.12

目录

- 背景
- 异常检测问题
  - 异常行为的种类
  - 常见的异常行为检测方法
  - 邮件系统中的异常行为检测
- 基于知识图谱的异常检测
  - 知识图谱的定义
  - 基于用户行为构建知识图谱
  - 基于图的异常检测模型
- 工作总结

背景

背景

- 近年来，随着我国各行业网络活动的日益增多，网络安全事件也频频发酵，这给校园网邮件管理系统带来了警示。如何防范账号盗用、垃圾邮件、信息泄露、病毒入侵等问题带来的安全风险，是需要我们持续探索的问题。
  - 2018年1月，某地方卫生系统出“内鬼”泄露50多万条新生婴儿和预产孕妇信息
  - 2018年2月，某员工私自转让公司权限给朋友，致使30余万条医生数据泄露
  - 2018年3月，某科技公司内鬼窃取500余万条个人信息，并在网上售卖
  - 2018年4月，北京某教育网站遭入侵，攻击者窃取7万余元
  - 2018年8月，某知名酒店集团5亿条用户数据泄露
  - ……



背景

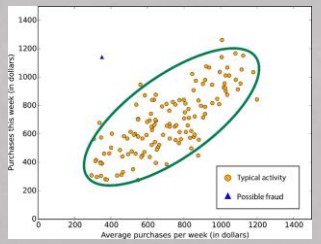
- 基于邮件系统的攻击行为案例
  - 2017年5月，WannaCry勒索病毒肆虐全球180个国家，钓鱼邮件是其重要的传播手段
  - 2017年12月，腾讯安全通报一起大范围钓鱼邮件攻击事件，近3万家中国企业受影响
  - 2018年6月，网络上出现了一个 14 亿邮箱密码泄露信息查询网站
  - 2018年9月，工信部监测发现近十万个互联网用户邮箱疑似被黑客控制
  - 2018年10月，国泰航空公布信息泄露事件之后，大量用户收到伪装地址的钓鱼邮件
- 基于邮件系统的攻击行为特点
  - 传播快、受众广、变种多
- 对于网络安全管理的启示
  - 引入新理念、新数据、新技术，适应网络安全新形态、新需求
  - 持续迭代更新网络威胁检测方法，提高网络安全管理能力

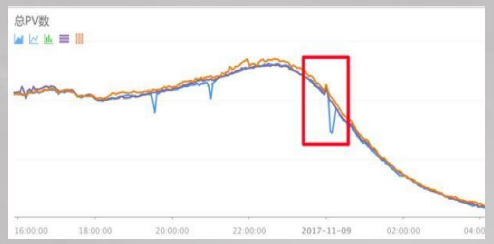
北京大學
PEKING UNIVERSITY

异常检测问题

异常行为检测问题

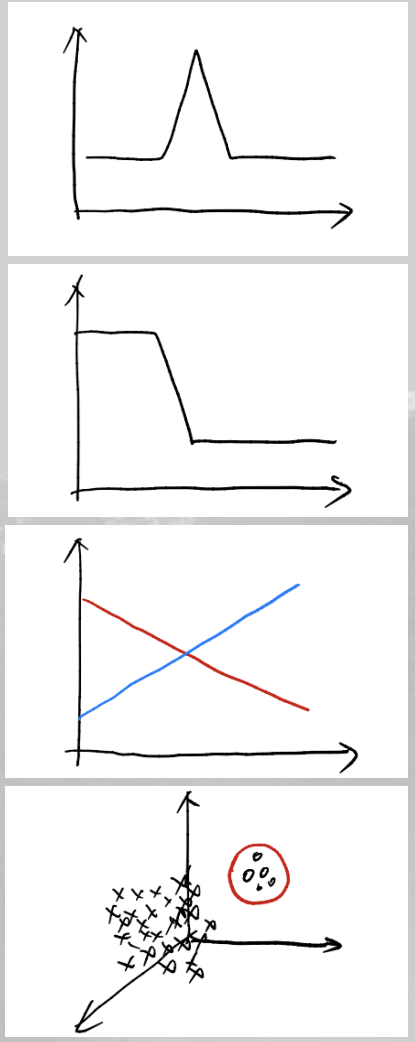
- **异常检测**
  - 是对稀有的事件、样本或观测值的识别，这些事件与大多数数据存在显著差异，从而形成异常；其应用包括疾病检测、金融欺诈检测、网络入侵检测等
- **异常行为**
  - 行人检测
    - 轮廓异常/姿态异常/异常集聚/…
  - 金融欺诈
    - 异常消费/风险操作/内幕交易/…
  - 网络入侵
    - 异常流量/异常登录/异常文件/…





邮件系统常见异常行为

- **时间序列数据**
  - 瞬时异常: 离群值 (outlier)
    - 宏观尺度上, 邮件系统短时间内用户登录行为激增/减
    - 微观尺度上, 用户登录/发信等行为远高于平均水平等
  - 持续异常: 时序变动 (temporal change)
    - 某段时间内用户数据持续异常等
  - 转化率异常: 水平变化 (level shift)
    - 用户不同指标间的转化率变化等
  - 群体异常: 群体性变化(group anomaly)
    - 大量用户多次短时间内进行相同操作，锁步行为
- **文本数据**
  - 邮件内容异常

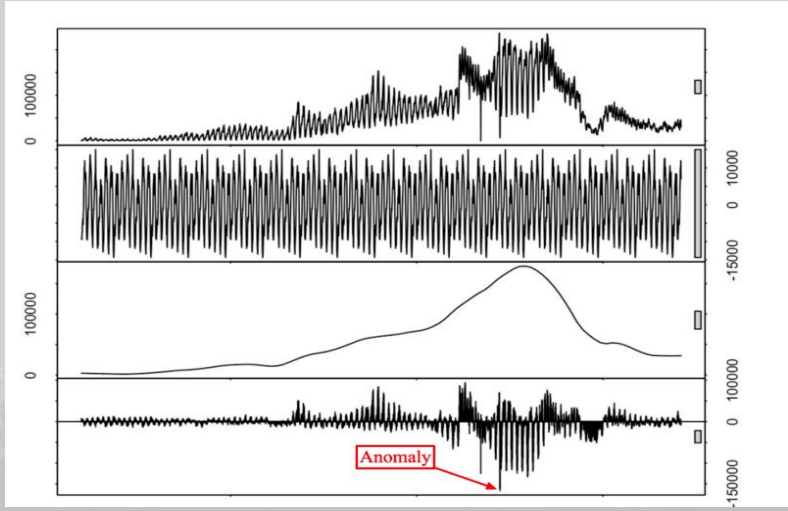


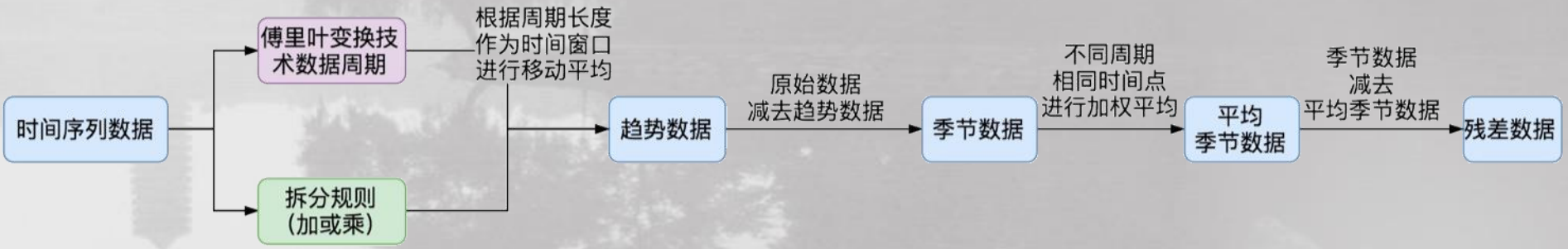
异常行为检测模型

- **基于规则**
  - 专家系统设计过滤规则
- **基于统计分析**
  - 3σ准则、箱型图、Grubbs检验
  - 时间序列建模: 移动平均、指数平滑、ARMA、ARIMA
- **基于学习的方法**
  - 有监督: 常见的分类模型，如 LR、SVM、RF、NN等方法
  - 半监督: 和无监督类似, one-class SVM、Auto-Encoder、GMM、Naïve Bayes等
  - 无监督: 基于统计分布、基于距离、基于密度、基于聚类和基于树的方法等

异常行为检测模型

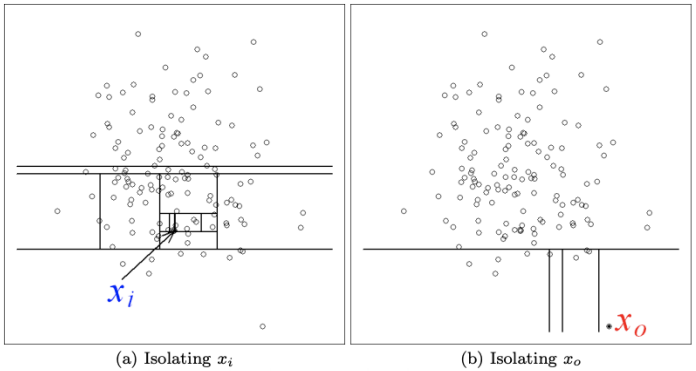
- **STL分解**
  - Seasonal + Trend + Loess
  - **选定周期窗口**
  - 从残差数据中实现异常检测
  - 假设残差服从正态分布，结合3σ准则筛选



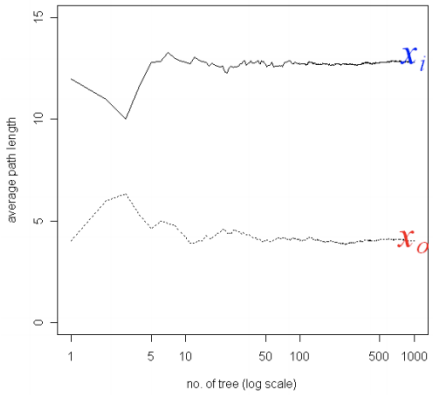


异常行为检测模型

- **孤立森林(Isolation Forest)**
  - 适用于**连续**数据的**无监督**异常检测方法
  - 异常被定义为“容易被孤立的离群点(more likely to be separated)”
  - 递归地随机分割数据集，直到所有的样本点都是孤立的，在这种随机分割的策略下，异常点通常具有较短的路径
  - 速度快，准确率高，效果好于常见的分类模型：one-class SVM/ LOF/Random Forests



(a) Isolating x_i (b) Isolating x_o



(c) Average path lengths converge

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation-based anomaly detection." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1 (2012): 3.

异常行为检测模型

- **变分自编码(VAE)**
  - 由变分网络(Variational net)和生成网络(Generative net)组成
  - 输入数据序列，经过变分网络可以得到一组隐变量，隐变量通过生成网络重构成原始输入
  - VAE学习到的是隐变量的分布，重构后的序列消除了噪声等影响
  - 可以使用重构误差来判断异常

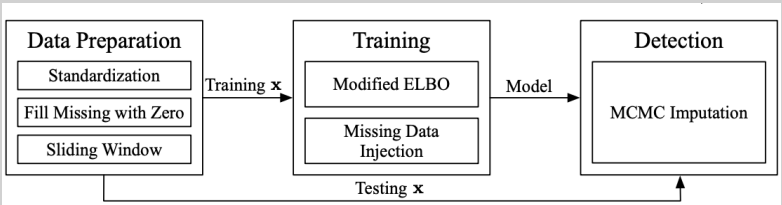
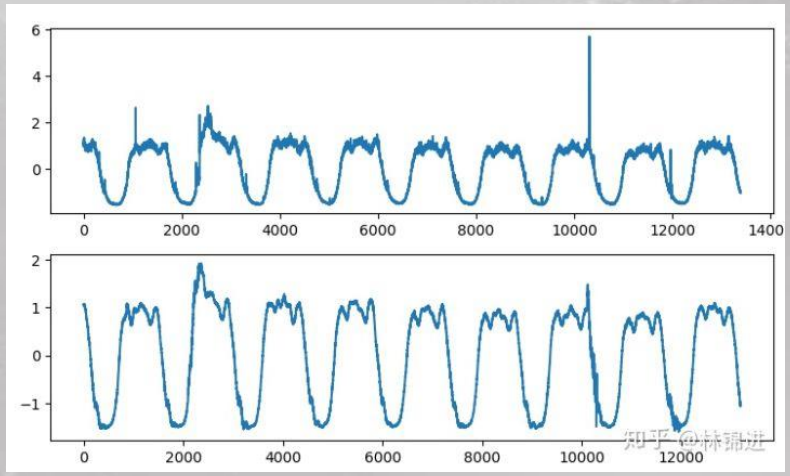


Figure 3: Overall architecture of *Donut*.



Xu, Haowen, et al. "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications." *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 2018.

邮件系统的异常检测

- **异常用户检测**
  - 异常用户检测
    - 异常使用频次检测
  - 异常行为检测
    - 异常使用习惯检测
    - 疑似僵尸账户检测
- **异常邮件检测**
  - 钓鱼/勒索邮件检测

北京大學
PEKING UNIVERSITY

基于知识图谱的异常检测模型

数据脱敏处理

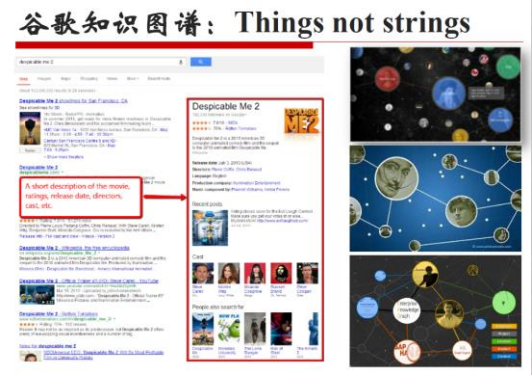
- **数据源**
  - 不涉及用户个人信息
- **敏感字段**
  - 对于学号、邮箱账号等字段进行加密
  - 通过哈希函数生成唯一id
- **敏感字段过滤、替换**
  - 剔除设备唯一识别码等

基于知识图谱的异常检测模型

- **基于用户画像构建知识图谱**
  - 引入多源跨域数据，建立用户画像
  - 根据用户画像挖掘用户关联关系
    - 建立用户之间的关系图谱
    - 建立用户与登录之间的行为图谱
- **基于知识图谱检测异常**
  - 基于OddBall算法挖掘异常用户
  - 基于一种改进的CopyCatch算法挖掘异常行为
  - 实验交叉验证

知识图谱的定义

- 知识图谱最早由谷歌公司于2012年5月提出，计划以此为基础构建下一代智能化搜索引擎，其关键技术包括从互联网的网页中抽取出实体及其属性信息，以及实体间的关系
- 国内业界有百度知心、搜狗知立方等商业应用；学术界有清华大学建立的第一个大规模中英文跨语言知识图谱XLore、中国科学院计算技术研究所的开放知识网络(OpenKN) 等



知识图谱的定义

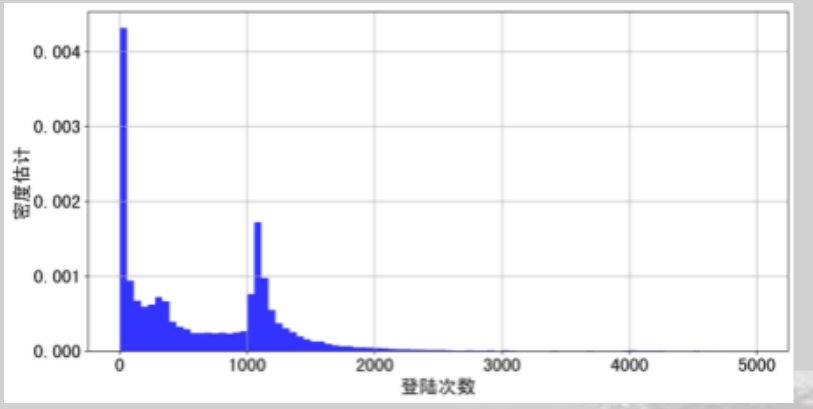
- **定义**
  - 知识图谱是结构化的语义知识库，以符号形式描述物理世界中的概念及其相互关系。其基本组成单位是<实体-关系-实体>三元组，以及实体极其相关属性-值对，实体间通过关系相互连结，构成网状的知识结构
- **要素**
  - **实体**：指的是具有可区别性且独立存在的某种事物。如某一个人、某一个城市、某一种植物等、某一种商品等等
  - **属性**：从一个实体指向它的属性值。不同的属性类型对应于不同类型属性的边。属性值主要指对象指定属性的值
  - **关系**： 形式化为一个函数，它把k个点映射到一个布尔值。在知识图谱上，关系是把k个图节点映射到一个布尔值的函数

知识图谱的关键技术

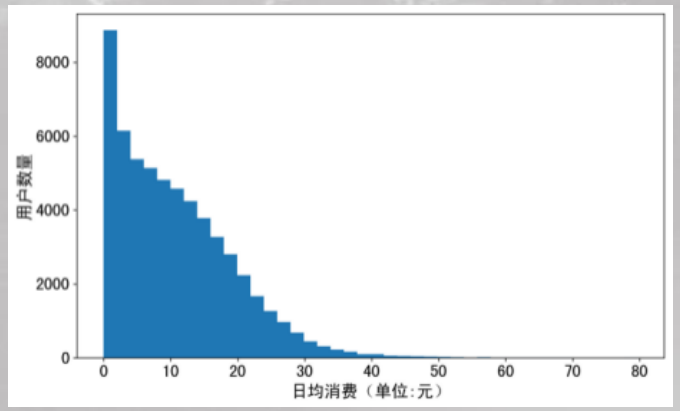
- **知识抽取**
  - 实体抽取、语义抽取、属性抽取、关系抽取
- **知识表示**
  - 距离模型、神经网络模型、矩阵分解模型
- **知识融合**
  - 消歧、去重、构建知识库等
- **知识推理**
  - 可满足性、分类、实例化
  - KBQA

用户行为分析

- **多源跨域数据融合**
  - 校园网邮件系统与一卡通系统共用校园卡号进行身份认证，因此可以融合邮箱登录行为、登录设备、校园卡消费记录等数据进行关联分析
- **用户行为挖掘**
  - 线上行为：邮箱使用习惯，包括日均登录频次、切换IP频次、登录设备类型等
  - 线下行为：校园卡消费习惯，包括日均餐饮消费、消费场所等



图：邮箱登录频次分布



图：校园卡日均消费金额分布

用户行为分析

- **属性抽取**
  - 属性包括: 入学年份、邮箱常用时段、常用登录地址、日均登录频次、设备持有量、餐饮日均消费、各餐次常用消费时段以及常用餐饮消费窗口等
- **特征选择**
  - 根据用户属性进行无监督聚类，并根据聚类簇赋予伪标签，从而选择有区分度的特征作为属性

用户行为分析

- **用户画像**
  - 通过定义用户的个体属性及常见的行为属性，实现了对用户特征的概括性描述，用于后续进一步挖掘
  - <用户-属性-值>构成基本的三元组，三元组之间以节点形式相互关联，形成图数据结构，可以用来更好的检索复杂的关联信息，从语义层面理解节点之间的关系

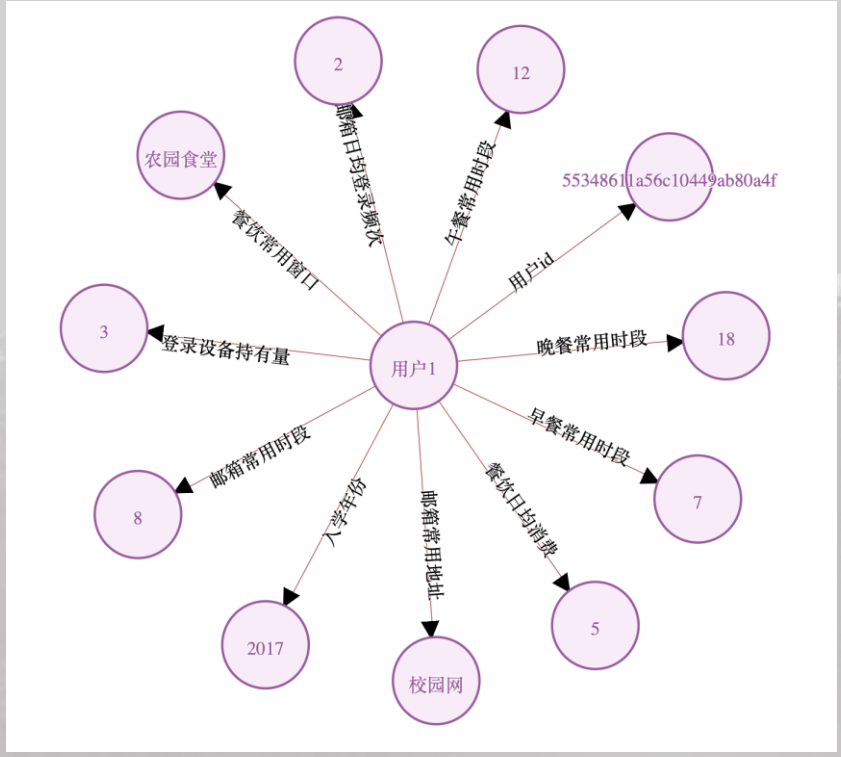
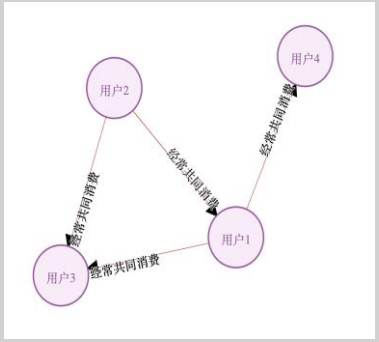


图: 某用户的用户画像

用户知识图谱

- **关系抽取**
  - 有别于从语料中进行关系抽取，我们根据分析目标直接定义实体间的关系
  - 如是否经常共同消费、是否曾在同一IP地址登录、生活规律是否相似等
  - 在异常检测中，关系的定义应有区分度
    - 消费行为中，我们绘制用户关系图的边权分布，结合阈值定义关系
    - 登录行为中，设置时间窗口阈值定义关系
  - 定义关系后，根据<实体-关系-实体>三元组建立相应的知识图谱



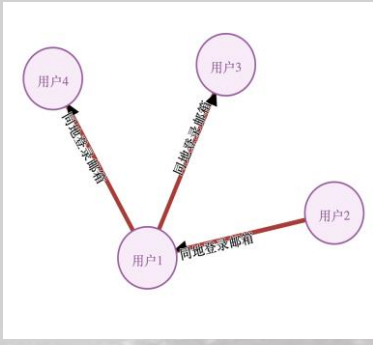


图: 基于不同关系建立的用户图谱

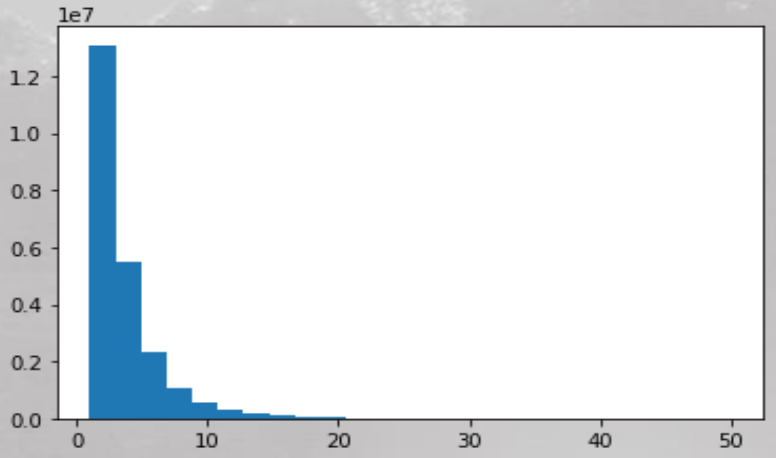


图: 用户关系图边权分布

用户知识图谱

- **线下消费关系图谱**
  - 用于衡量用户之间的线下关系是否紧密，从而发现异常线下行为特征

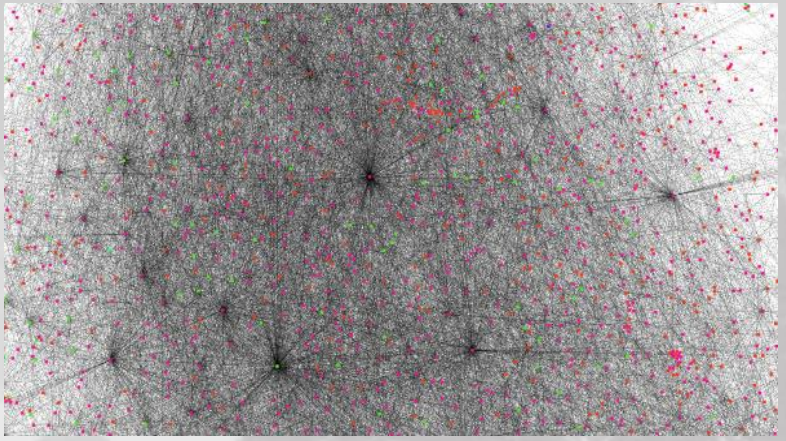


图: 基于9000余名用户建立的关系图谱

- **线上登录行为图谱**
  - 用于衡量用户线上行为是否与群体一致，从而发现异常登录行为模式

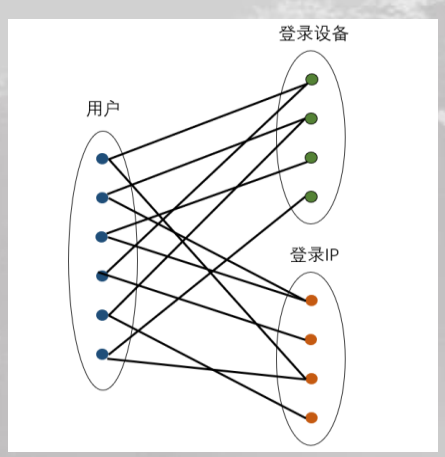


图: 用户登录行为二部图

基于知识图谱的异常检测

- **异常用户检测: 线下行为**
  - 挖掘线下行为中的异常用户
  - 线下消费的几种异常模式:
    - 星型, 重近邻, 主导边, 社团
- **OddBall算法**
  - 针对实体节点建立ego-net
  - 针对三种异常模式绘制不同的关系图
  - out-line: 计算点与拟合直线的偏离程度
  - out-lof: 计算点与聚类的偏离程度

$$out\text{-}score(i) = out\text{-}line(i) + out\text{-}lof(i)$$

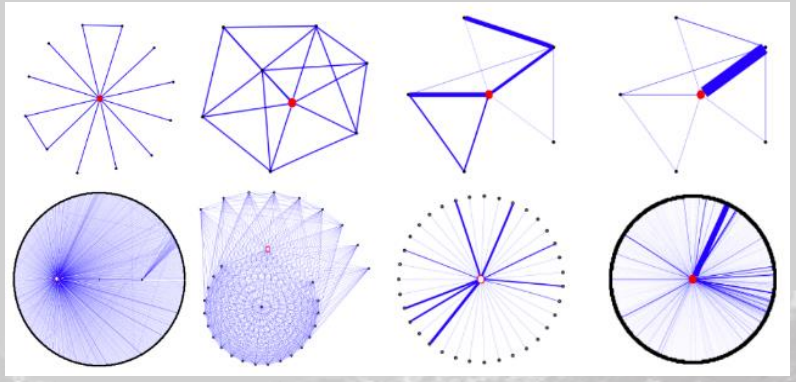
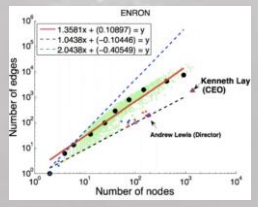
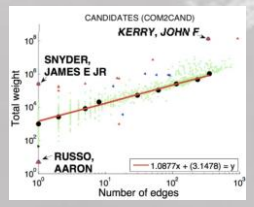


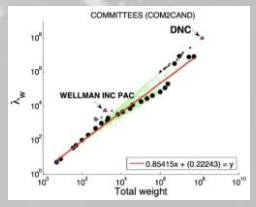
图: 几种异常模式



Clique-Star: E v.s. N



Heavy Vicinity: W v.s. E



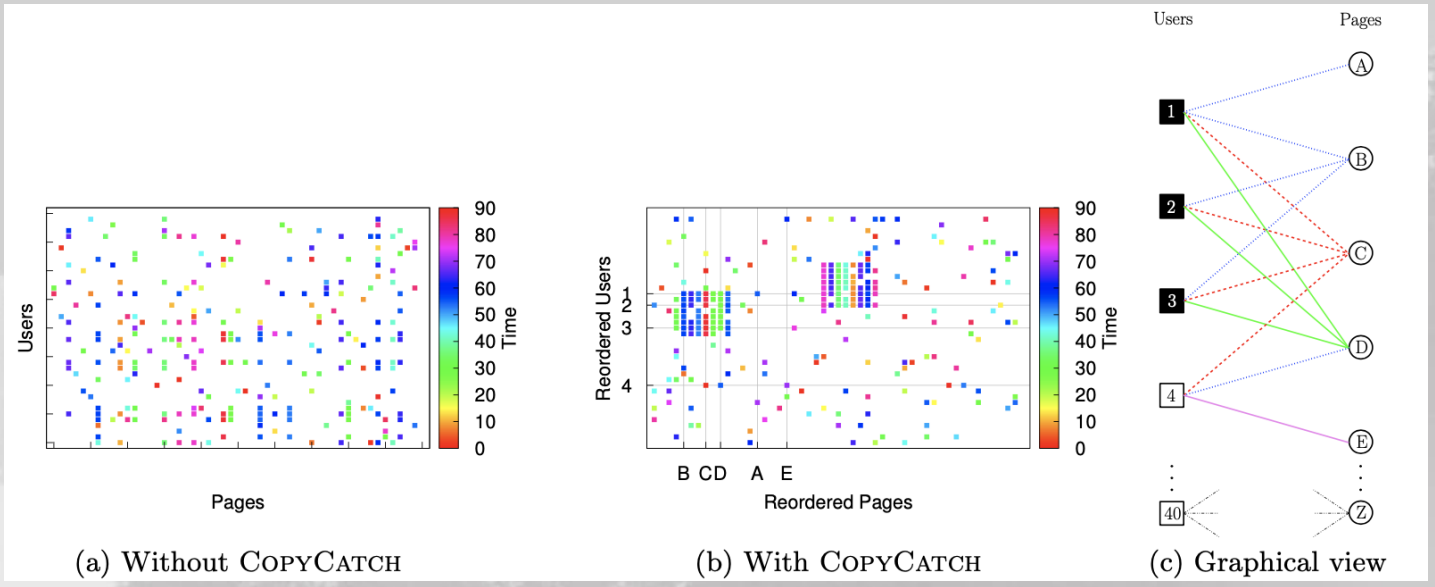
DominantPair: λ v.s. W

Ni: i的度数　　Ei: i的边数
λw,i: i ego-net 邻接矩阵的主特征值
Wi: i ego-net的总权重

Akoglu, Leman, Mary McGlohon, and Christos Faloutsos. "Oddball: Spotting anomalies in weighted graphs." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 2010.

基于知识图谱的异常检测

- **异常用户检测: 线上行为**
  - 个体异常用户
  - **群体异常用户**



(a) Without CopyCatch (b) With CopyCatch (c) Graphical view

Beutel, Alex, et al. "Copycatch: stopping group attacks by spotting lockstep behavior in social networks." *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

基于知识图谱的异常检测

- **CopyCatch算法**
  - 锁步问题
    - 假设有N个用户$U = \{i\}_{i=1}^{N}$以及M个日期$P = \{j\}_{j=1}^{M}$, 则分别构造
      - 指示矩阵$I = \{I_{ij}\}$, 如果用户i在日期j有过登录行为, 则I_{ij} =1;
      - 数据矩阵$L = \{L_{ij}\}$, 如果用户i在日期j的时刻t_{ij}有过登录, 则$L_{ij} = t_{ij}$;
    - 锁步问题可以描述为:定义时间关联二分核为$[n, m, \Delta t, \rho]$，假设存在一系列用户$U' \subseteq U$ 和一系列日期 $P' \subseteq P$, 如果存在 $u_i' \subseteq U'$ 和 $p_i' \subseteq P'$ 满足如下条件，则可以推断出u_i'和p_i'存在锁步行为:

$$|U'| \geq n$$
$$|P'| \geq m$$
$$|p_i'| \geq \rho |P'| \quad (\forall i \in U')$$
$$(i,j) \in \varepsilon \quad (\forall i \in U', \forall j \in P')$$
$$\exists t_j \in R_{s.t.} \ |t_j - L_{i,j}| \leq \Delta t \quad (\forall i \in U', \forall j \in P')$$

基于知识图谱的异常检测

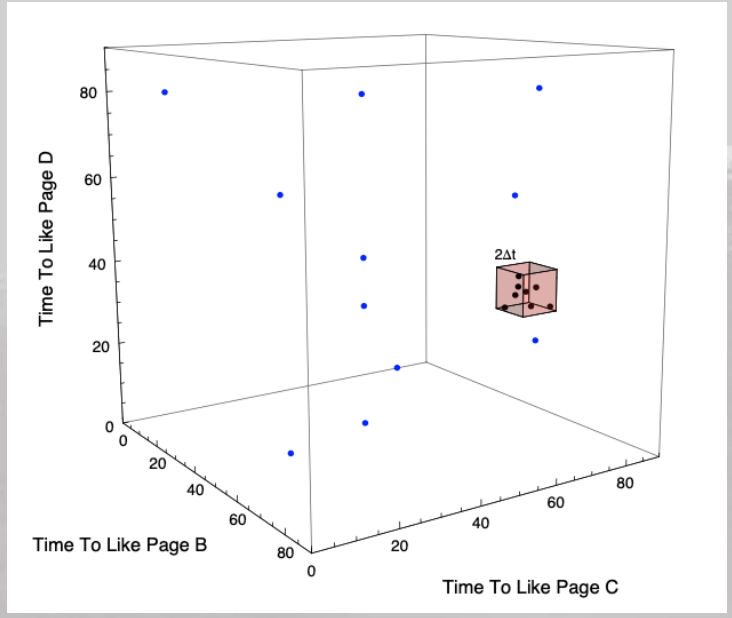
- **CopyCatch算法**
  - 锁步问题的最优化

$$\max_{c,P':|P'|=m} \sum_i q(\mathrm{L}_{i,*}|c,P')$$

where

$$q(u|c,P') = \begin{cases} \sigma \ \ if\ \sigma = \sum_{j\in P'} \mathrm{I}_{ij}\phi(c_j,u_j) \geq \rho m \\ 0 \ \ otherwise \end{cases}$$

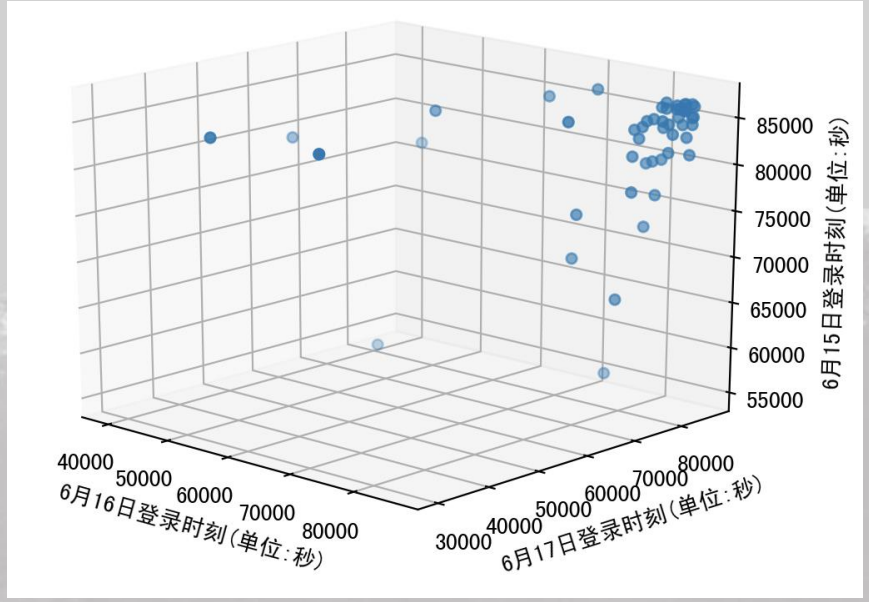
$$\phi(t_c,t_u) = \begin{cases} 1 & if\ |t_c - t_u| \leq \Delta t \\ 0 & otherwise \end{cases}$$



  - 其中L代表用户登录设备的数据矩阵，c是欺诈用户发生欺诈行为的中心向量，P'是伪造登录设备集合，ρ为松弛因子

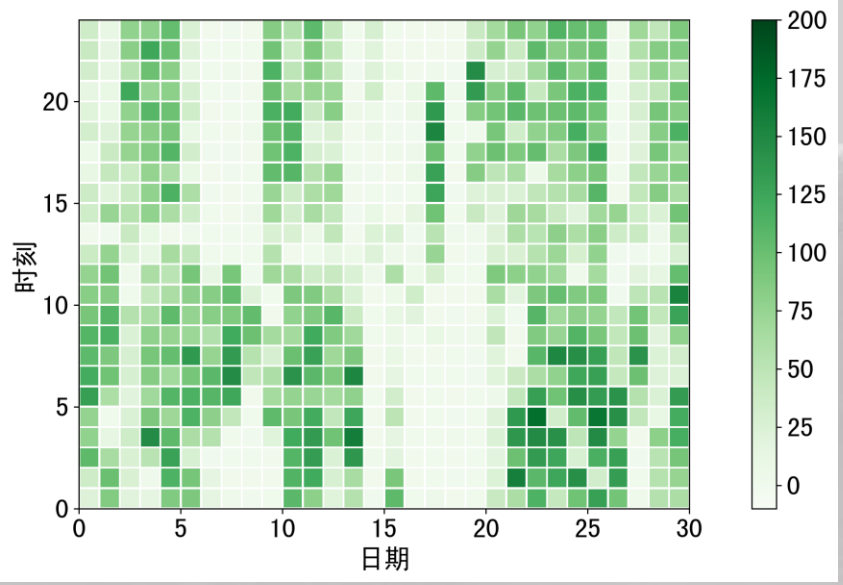
实验结果

- **线上异常行为检测**
  - 在不同日期，同一批账号在较短时间内同时登录，具有“锁步”特征



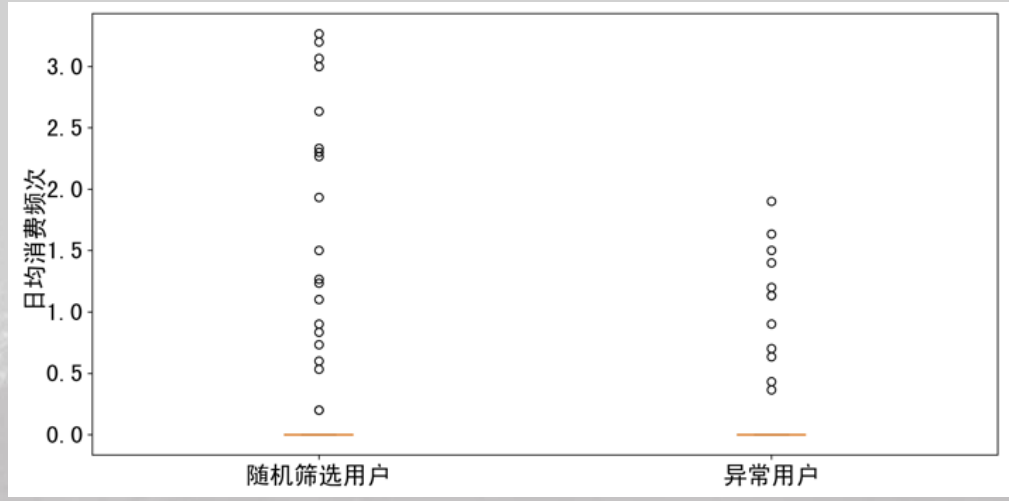
实验结果

- **线上异常行为检测**
  - 选择其中一个异常用户绘制登录热力图
  - 该用户频繁登录邮箱，几乎每天都有上百条登录记录，且登录时段集中于凌晨0-5点之间，体现出较明显的异常行为特点



实验结果

- **线上异常行为检测**
  - 利用线下消费行为进行验证
  - 可以发现算法检测的异常用户线下活动的频次也低于正常用户
  - 结合该类异常用户经常“锁步”登录以及线上行为异常活跃等特点，我们推测这类用户可能为信息泄露的离校毕业生等，其账户信息被攻击者盗用进行批量登录尝试。



实验结果

- **线下异常用户检测**

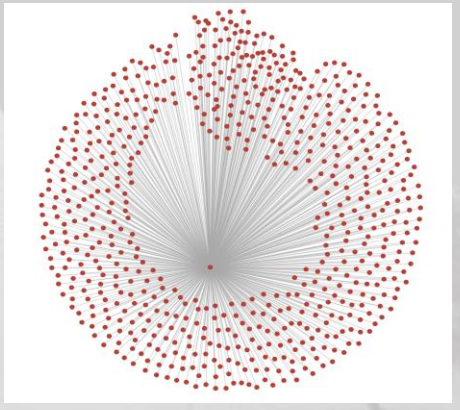


图: 星型

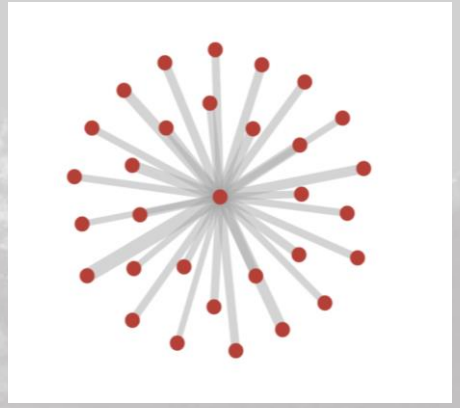


图: 重近邻

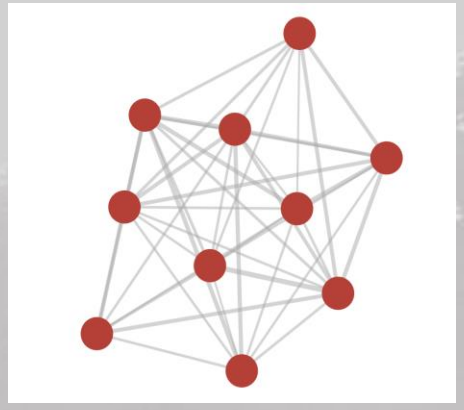


图: 社团结构

实验结果

- **线下异常用户检测**
  - 验证异常用户线上行为

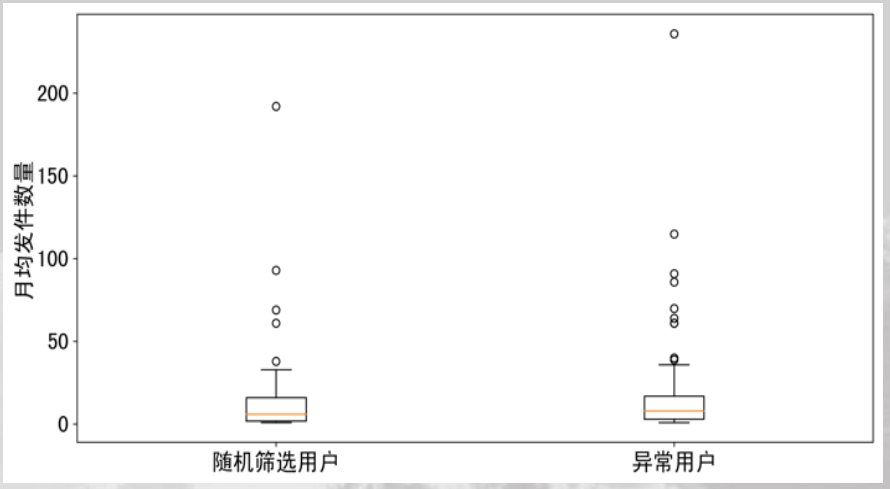


图: 异常用户月均发件数量

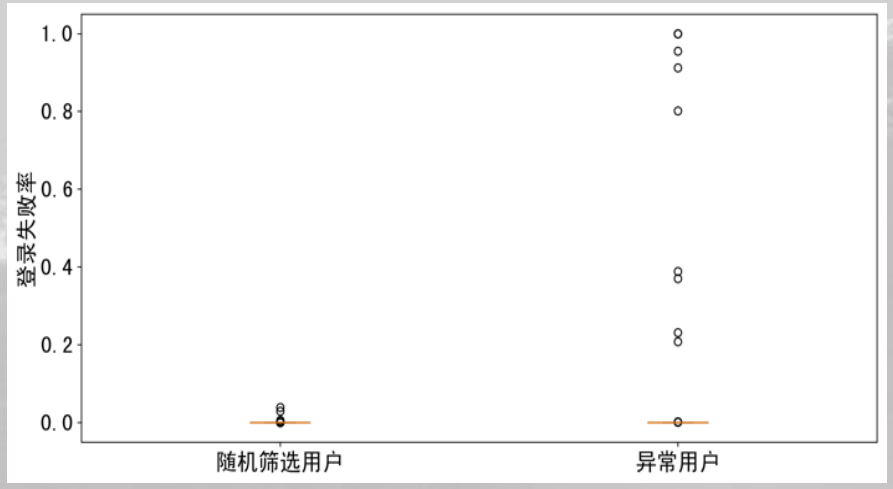


图: 异常用户登录失败率

总结

总结

- **知识图谱的设计**
  - 基于用户画像构建了用户行为知识图谱
- **基于知识图谱的异常检测算法的实践**
  - 采用OddBall和CopyCatch算法实现线上和线下异常行为的检测
  - 在群体异常行为检测部分，结合邮件数据集特有的结构和特征，对算法进行了改进，有效的检测了批量登录尝试这一异常行为
- **利用线下数据辅助线上异常行为的检测**
  - 在用户消费行为知识图谱上检测到的线下异常用户，其线上行为往往也存在异常

未来工作

- **设计更全面的知识图谱**
  - 引入更多的数据源, 包括但不限于AP数据, 网关登录系统数据, 收发信数据等
  - 构建更多关系的知识图谱
- **实现基于知识图谱的推理**
  - 引入知识图谱中包含的语义信息, 实现基于图谱的推理检测
- **增加更多实验**
  - 标注异常用户数据集, 与经典异常检测方法进行性能测试

感谢批评指正！