You are How You Click:

Clickstream Analysis for Sybil Detection

Lyu Jiuyang, Jan 3rd, 2022

介绍

目标： 虚假身份账户检测。

动机： 现有的条形码验证机制和基于图的检测机制未被证实有效。

核心： 通过划分捕获点击流序列之间距离的相似性图，将相似的用户点击流分组。

相关工作： 虚假用户难以与真实用户建立关系，将其社交网络形成强连通子图，用图论解决。但是有效性不明确。

名词：

- 点击流（clickstream）：在线用户在每次 Web 浏览Session期间生成的点击事件的踪迹。本文特指用户向网站发出的 HTTP 请求序列。
- 虚假身份账户（Sybils）

贡献：

- 作者[首个]提出了多种点击流模型进行用户点击模式的表征，并将其映射到相似性图中。
- 作者开发了一个检测系统，效果良好。

方法论

初步分析

Session级特征

Sybils 会花费大量点击发送好友请求和浏览个人资料。

Category	Description	Sybil Clks # (K)	Sybil Clks %	Nrml Clks # (K)	Nrml Clks %
Friending	Send request	417	**41**	16	0
	Accept invitation	20	2	13	0
	Invite from guide	16	2	0	0
Photo	Visit photo	242	**24**	4,432	**76**
	Visit album	25	2	330	**6**
Profile	Visit profiles	160	**16**	214	4
Share	Share content	27	3	258	**4**
Message	Send IM	20	2	99	2
Blog	Visit/reply blog	12	1	103	2
Notification	Check notification	8	1	136	2

Table 2: Clicks from normal users and Sybils on different Renren activities. # of clicks are presented in thousands. Activities with <1% of clicks are omitted for brevity.

Session期间的活动

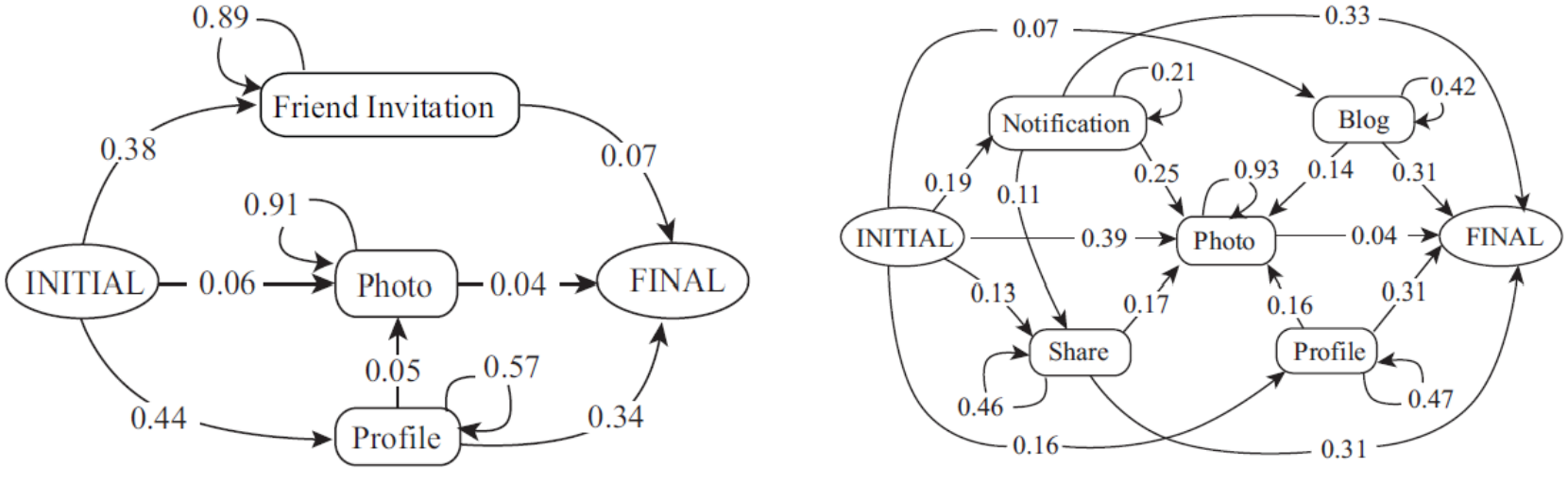
观察8个类别共55项活动

- 加好友：包括发送好友请求、接受或拒绝这些请求以及取消好友关系。
- 照片：包括上传照片、整理相册、标记朋友、浏览朋友的照片以及对照片发表评论。
- 个人资料：浏览用户个人资料。
- 分享：指用户在他们的主页上展示超链接。
- 消息：包括状态更新和即时消息 。
- 博客：包括撰写博客、浏览博客文章以及在博客上发表评论。

- 通知：指点击人人网的通知机制，提醒用户对其内容发表评论或点赞。
- 赞/踩

点击活动的转换

建立**马尔可夫链**：每个类别都是一个状态，边代表状态之间的转换。可以看出Sybils和真实用户在行为上有明显区别，他们更多只进行加好友、发垃圾照片和浏览个人信息。



(a) State transitions for a Sybil account. (b) State transitions for a real user.

Figure 7: Categories and transition probabilities in the clickstream models of Sybils and normal users.

SVM

使用的Features：

- Session特征：从Session中提取的 4 个特征：每个会话的平均点击次数、平均会话长度、点击之间的平均到达间隔时间，以及每天的平均会话数。
- 点击功能：对于每个用户，我们使用每个类别的点击百分比作为一个特征。

以上方法都为有监督学习，接下来进行无监督学习。

点击流建模和集群

点击流模型

点击序列模型

$$S = (s_1 s_2 \dots s_i \dots s_n),$$

基于时间的模型

$$[t_1, t_2, t_3, \dots, t_n]$$

二者结合

$$a(t_1)c(t_2)a(t_3)d(t_4)b$$

a, b, c, d为点击类型，t_i为间隔时间。可以按照细粒度（55）或者粗粒度（8）进行类型的对应。

计算序列相似度

使用 `k-gram` (长度为k的子序列)

$$T_k(S) = \{k-gram | k-gram = (s_i s_{i+1} \dots s_{i+k-1}), i \in [1, n+1-k]\}.$$

公共子序列

使用类似 `Jaccard Coefficient` 作为两个序列相似度的表示。

$$D_k(S_1, S_2) = 1 - \frac{|T_k(S_1) \cap T_k(S_2)|}{|T_k(S_1) \cup T_k(S_2)|}$$

公共子序列次数

$$D(S_1, S_2) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{n} (c_{1j} - c_{2j})^2}$$

基于分布的方法

之前的方法无法处理基于时间的模型，对于连续值序列 S_1, S_2，通过 双样本 KolmogorovSmirnov 检验（K-S 检验）来计算距离，这个方法对两个样本的经验累积分布函数 (CDF) 的位置和形状的差异很敏感。

$$D(S_1, S_2) = \sup_t |F_{n,1}(t) - F_{n',2}(t)|$$

其中，$F_{n,i}(t)$是序列S_i的CDF值， $\sup$ 指序列能到达的最大值

Model	Distance Metrics
Click Sequence Model	*unigram*, *unigram+count*, *10gram*, *10gram+count*
Time-based Model	*K-S* test
Hybrid Model	*5gram*, *5gram+count*

Table 4: Summary of distance functions.

序列集群

下一步需要将具有相似点击流的用户聚集在一起，作者构建并划分了一个序列相似度图，每个用户的点击流由一个节点表示。每对节点之间边的权重是序列之间的相似性距离。划分的过程为最小化断边的总权重，这样相似活动（权重高）的用户将会被划分到一个集群。

本文作者使用METIS算法，将图划分为K个集群，K为超参数。

效果分析

其中，`*gram` 代表使用1、2方法计算距离时 `k-gram` 的长度，`unigram` 代表1，

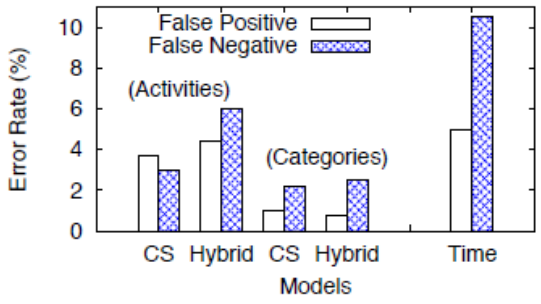


Figure 8: Error rate of three models.

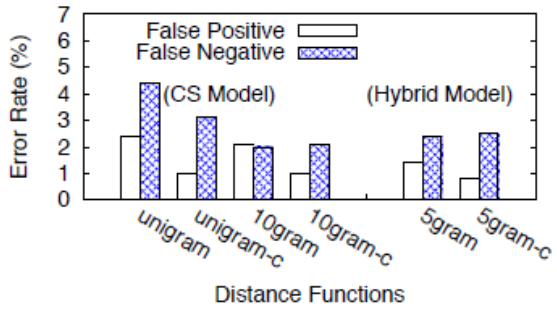


Figure 9: Error rate using different distance functions.

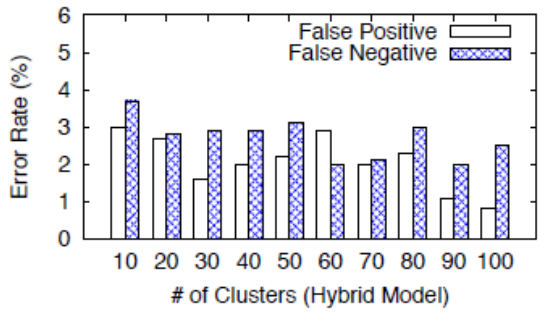


Figure 10: Impact of number of clusters (K).

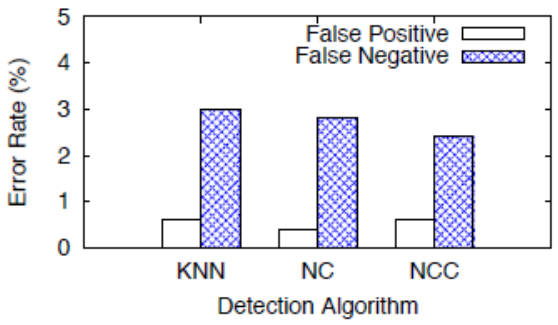
- CS 和 Hybrid 模型明显优于基于时间的模型
- 基于类别的点击编码优于按活动编码
- 10gram+count 和 5gram+count 分别是CS和Hybrid模型的最佳距离函数，K=100最佳

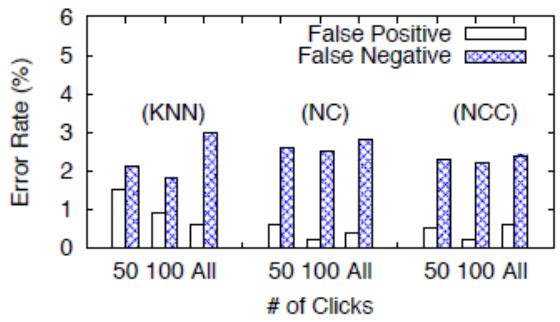
Incremental Sybil Detection （设计系统）

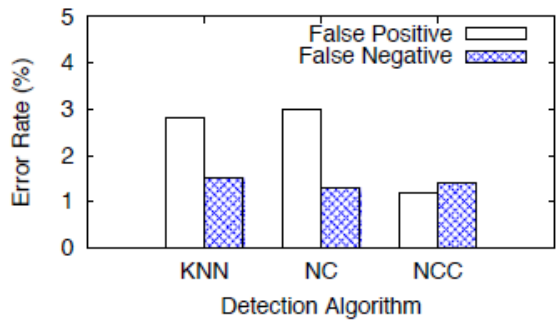
将新序列“重新聚类”到现有图中，作者研究了三种算法：

- KNN
- 最近聚类 Nearest Cluster
- Nearest Cluster-Center (NCC)，计算密集程度较低的 Nearest Cluster，只需要计算未分类序列到每个现有集群中心的距离

预计算中心序列：与同一簇中所有其他序列距离最短的序列







无监督学习

需要少量已知真实用户的点击流作为“种子”，为他们所在的集群着色。这些种子可以根据需要手动验证。 我们将包含种子序列的所有集群着色为“正常”，而未着色的集群被假定为“Sybil”。

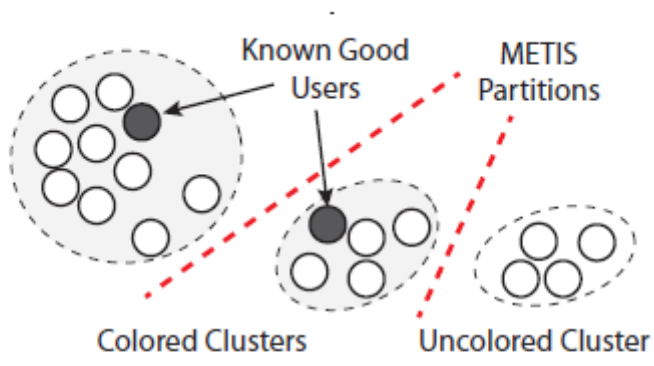


Figure 14: Unsupervised clustering with coloring.

经验证，需要较少的seed就可以为所有的集群“着色”。seed set不需要随着时间的推移而彻底改变。改变不同的seed数量和数据集正负例比例影响如图。

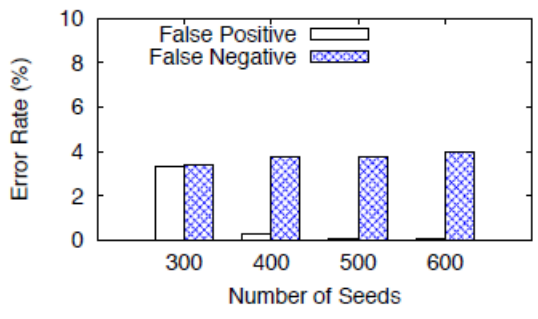


Figure 17: Detection accuracy versus number of seeds.

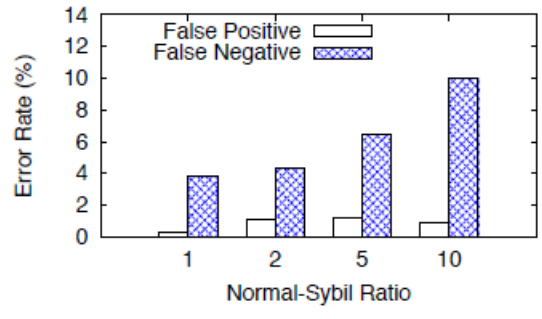


Figure 18: Detection accuracy versus Normal-Sybil ratio.

持续对抗分析（7.2 Limits of Sybil Detection）

其他

- 点击流模型的应用
- 本文的一大亮点是无监督学习（为了少数据样本），自监督学习也是可以考虑使用的一种方法
- METIS - Serial Graph Partitioning and Fill-reducing Matrix Ordering