DOI:10.7544/issn1000-1239.2021.20210461
58(11): 2319-2332, 2021

微信恶意账号检测研究

杨 征[1] 殷其雷[1] 李浩然[1] 苗园莉[1] 元 东[1] 王 骞[2] 沈 超[3] 李 琦[1]
[1](清华大学网络科学与网络空间研究院 北京 100084)
[2](武汉大学国家网络安全学院 武汉 430072)
[3](西安交通大学网络空间安全学院 西安 710049)
(yz17@mails.tsinghua.edu.cn)

Study of Wechat Sybil Detection

Yang Zheng[1], Yin Qilei[1], Li Haoran[1], Miao Yuanli[1], Yuan Dong[1], Wang Qian[2], Shen Chao[3], and Li Qi[1]
[1](*Institute for Network Sciences and Cyberspace*, *Tsinghua University*, *Beijing* 100084)
[2](*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)
[3](*School of Cyber Science and Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049)

Abstract Online social networks (OSNs) are efficient platforms for information dissemination and facilitate our daily life. The value of OSN accounts increases with the popularity of OSNs. In order to obtain profits illegally, attackers leverage OSNs to construct various attacks such as fraud and gambling. A number of solutions have been proposed to protect users' security, which mainly focuses on detecting malicious accounts (or Sybils) by analyzing user behavior or the propagation of user relations. Unfortunately, it usually takes much time to collect enough data to perform malicious account detection. Attackers can perform different kinds of attacks during the data collection phase. To detect Sybils efficiently, we propose a new approach that leverages account registration attributes to detect Sybils. First, we analyze the existing detection methods in sybil detection. Then, we analyze the registration data of WeChat. We analyze and compare the distribution of Sybils and benign accounts in different registration attributes, and find that Sybils are prone to cluster with some registration attributes. According to these statistics, we extract two kinds of features from different attributes, i. e., synchronization-based features and anomaly-based features, and calculate the similarity of two accounts based on those features. The accounts that have high similarity are more likely to be malicious. Finally, we build a graph upon accounts having a high similarity to cluster malicious users. We calculate a malicious score for each user to infer whether it is a Sybil. We prototype our approach, and the experimental results with real WeChat show that our approach can achieve 96% precision and 60% recall.

Key words online social network; clustering; sybil detection; account registration characteristics; statistical analysis

收稿日期:2021-05-11;**修回日期**:2021-08-13

基金项目:国家重点研发计划项目(2018YFB1800304);国家自然科学基金项目(61572278,U20B2049,61822207,61822309,61773310,U1736205,62132011);北京信息科学与技术国家研究中心项目(BNR2020RC0101);陕西省重点产业创新链项目(2021ZDLGY01-02)
This work was supported by the National Key Research and Development Program of China (2018YFB1800304), the National Natural Science Foundation of China (61572278, U20B2049, 61822207, 61822309, 61773310, U1736205, 62132011), the Project of BNRist (BNR2020RC0101), and the Shaanxi Province Key Industry Innovation Program (2021ZDLGY01-02).

通信作者:李琦(qli01@tsinghua.edu.cn)

摘 要 社交网络是一个有效的信息传播平台,使得人们的生活更加便捷.同时,在线社交网络也不断提高了社交网络账号的价值.然而,为了获取非法利益,犯罪团伙会利用社交网络平台隐秘地开展各种诈骗、赌博等犯罪活动.为了保护用户的社交安全,各种基于用户行为、关系传播的恶意账号检测方法被提出.此类方法需要积累足够的用户数据才能进行恶意检测,利用这个时间差,犯罪团伙可以开展大量的犯罪活动.首先系统分析了现有恶意账号检测工作.为克服现有方法的缺点而更快地检测恶意账号,设计了一种基于账号注册属性的恶意账号检测方法.方法首先通过分析恶意账号和正常账号在不同属性值上的分布,设计并提取了账号的相似性特征和异常特征;然后基于此计算两两账号的相似度构图以聚类挖掘恶意注册团体,从而有效实现注册阶段的恶意账号检测.

关键词 在线社交网络;聚类;恶意账号检测;账号注册属性;统计分析

中图法分类号 TP393

随着移动互联网的发展,移动社交网络迅速地成为了主流的社交平台,在人们的生活和工作中占据了重要的地位.除了传统的对话交流,移动社交网络平台还为用户提供了更加多样化的服务,例如阅读、购物等.众多的用户共同编织了一张复杂的社交网络,他们的一举一动都在影响着这个网络的发展.与此同时,一些不法分子也在社交网络上活跃着,他们使用批量注册的账号,在社交平台上进行各种恶意活动,例如赌博、刷阅读量、引导用户流等,以此牟取不法利益.

为了防止恶意账号危害社交网络生态环境,许多恶意账号检测模型被提出.例如基于用户发送的内容(如文字、图片、链接等)、用户行为(如点赞、上传、关注等),来训练检测模型[1-6];或者根据用户间的社交关系(如好友、关注、订阅等)构建图模型,来挖掘恶意团体[7-10].然而,无论是用户发送的内容、用户的行为,还是用户间的社交关系,都需要一定的时间来收集和积累.在这段期间里,恶意账号已经可以开展大量的恶意活动,对正常用户造成影响.另一方面,当今黑色产业链已大规模地采用自动化账户注册技术来批量获取社交网络账号,以确保其恶意活动可持续、影响范围广且获得的不法利益高.本文首先具体介绍了现有恶意账号检测工作并分析了它们的优缺点.为克服它们的局限性,更加快速而有效地应对此类黑产威胁,减少恶意账号造成的危害,并尽早检测出此类由黑产链批量注册的恶意账号,本文提出了基于社交网络账号注册属性的恶意账号检测方法.本文工作旨在仅基于账号注册属性实现对恶意账号的有效检测,在注册阶段即遏止恶意账号的进一步活动.

微信是现今中国最大的移动互联网社交网络平台,每天的账号注册量可达百万级别.在与微信平台所有方——腾讯公司——达成深度合作的基础上,本文对其提供的 2017 年部分微信注册账号数据进行了深入统计分析,发现在某些时段上恶意账号占全部新注册账号的比例可高达 50%,但由于涉及商业隐私,我们无法给出具体的数值.此外,还发现恶意账号会使用某些相同的注册属性,如某个恶意团体注册的账号具有相同的 IP 前缀和手机号码前缀.这是由于恶意团体注册账号时受时间、人力、设备等资源所限,使用机器批量注册造成的.此外,恶意账号的某些注册属性会存在异常,如注册国家与用户填写的国家不一致、注册 IP 所在省份与手机号所在省份不一致等.这可能是因为这些属性是黑色产业链的批量注册工具随机生成的,未考虑正常账号属性间的内在联系.针对恶意账号具有的相似性特征和异常特征,本文使用了带权重的无向非连通图的账号聚类算法,用图中的点代表账号,边代表账号间的相似关系,图中的连通分量便能体现账号间的群组关系,从而能够快速地将大量注册账号聚集成若干个群组.注意到每一个连通分量都是一张有权重的无向连通子图,本文提出了基于带权重的无向连通图的恶意检测算法,为每个账号计算出恶意分数来衡量其恶意程度.恶意分数高的账号将最终被判定为恶意账号.

本文的贡献主要有 3 个方面:

1) 系统分析了近年来社交网络恶意账号检测的研究工作,包括基于账号属性特征的检测模型与基于账号间关系的检测模型;

2) 对微信账号的注册数据进行大规模地统计分析,系统总结了恶意账号具有相似与异常的注册属性模式;

3) 设计了一种基于注册属性的无监督恶意账号检测方法,可在账号注册阶段实现对大规模恶意

账号的快速有效检测,且该方法无需提供标签数据作为训练集.

1　社交网络恶意账号检测工作

目前已经有很多在社交网络上检测恶意账号的工作.这些工作主要可以分为两大类模型:1)基于账号属性特征的检测模型;2)基于账号间关系的检测模型.

基于账号属性特征的检测模型通常将检测恶意账号视为一个机器学习领域的二分类问题[1-6,11-18].根据每个账号的自身属性如发布的内容(如微博、Twitter中的URL等)、操作行为(如点击流、关注、订阅等)、注册信息(如IP地址、User Agent等)来提取相关特征,然后使用提取的特征和有标签的数据来训练有监督的机器学习模型.如Almaatouq等人通过分析Twitter用户的行为、消息内容以及用户的个人画像信息,提取特征区分正常账号和不同类型的恶意账号[1];Egele等人基于用户行为建立用户画像,通过用户行为发生突变的异常现象,结合恶意账号活动具有相似性的特点进行检测[2];Freeman等人利用账号登录的数据,从概率统计分析的角度检测用户是否为恶意账号[3];Badri等人通过用户的点赞行为提取特征区分恶意账号[4];Wang等人通过对用户点击行为数据划分会话,在此基础上进行会话和点击行为流2个层面的特征提取[5];Thomas等人基于账号注册时的昵称、邮箱名等信息提取恶意注册模式,并结合用户注册的行为流、用户代理等信息检测恶意账号[18].

然而真实社交网络的用户数量是巨大的.以微信为例,每天新注册的用户数量在百万级,且给如此多的数据标注标签是不现实的.这导致基于账号自身属性和有监督机器学习算法的检测模型难以在大规模的社交网络中进行实际使用.另一方面,使用有监督方法检测恶意账号的方法鲁棒性难以得到保障,一旦检测使用的模型或特征提取的方式不慎泄露,恶意用户可以有针对性地修改自己的行为模式和个人信息,使得模型分类效率大幅下降从而逃避检测.而本文提出的检测模型是无监督的,不依赖标签,使得我们的检测模型更有实用价值.

基于账号间关系的检测模型通常使用图模型来刻画账号间关系,故可被称为基于图拓扑的检测模型.其通常以用户为点,用户之间的关系为边建图,利用图的拓扑结构发现恶意账号[7-10,19-39].常见的用户间的关系有:好友关系、关注与被关注、订阅与被订阅、相同或相似的行为、使用相同的设备或资源等.比如Jiang等人根据分析账号在社交网络上的行为来建立一张账号行为关系图,通过图拓扑分析账号的同步性和异常性,进而发现恶意账号[7];或者利用图的信息传播特点挖掘恶意账号,例如通过受害账号及其社交关系拓扑找出恶意账号[8],通过部分账号标签和账号间社交关系构成的社交网络图推导其他账号的标签[9],或计算信任关系在图中的传播从而发现信任度低的团体[10].

然而,这些已有的基于图拓扑的社交网络恶意账号检测方法都依赖用户在社交网络上产生足够的行为或建立足够多的社交关系.这就意味着只有当恶意账号在社交网络上活跃了一段时间,比如几天、几周甚至几个月等,这些图模型检测方法才能够将它们检测出来.而本文提出的检测模型,是利用用户的注册信息来进行恶意账号检测的,检测的时间点是账号注册当天,从而极大压缩恶意账号在社交网络上的存活时间.

Thomas等人的工作与本文方法在检测算法设计上较为相似,不过他们检测恶意账号时使用了较多只适用于Twitter网站的数据与特征,如注册流、User Agents、表单提交时间等[18],且运用了各账号在网页上的具体操作与交互数据.而我们的工作仅使用了更加通用且仅提取自账号注册阶段的特征,如IP地址、手机号等,从而可仅基于账号注册信息进行有效快速的判断.Yuan等人提出了Ianus方法同样可在注册阶段检测恶意账号,但是该方法需要账号的标签数据来进行训练和调整[40].而本文方法不依赖标签数据,只通过对比和度量各账号间注册属性相似性,即可构建账号相似连通图并挖掘由疑似恶意账号所组成的连通分量,属于无监督类方法,故适用性更佳.如表1所示,我们详细地列出了各检测方法所运用的信息与本文方法的差异.

Table 1　Comparison of Sybil Detection Methods on Social Network

表1　社交网络恶意账号检测方法对比

类别	模型	内容或行为		社交关系		检测实时性
		图拓扑	特征	图拓扑	特征	
有监督	文献[8]	×	×	√	×	×
	文献[9]	×	×	√	×	×
	文献[10]	×	×	√	×	×
	文献[20]	×	×	√	×	×
	文献[21]	×	×	√	×	×
	文献[22]	×	×	√	×	×

续表 1

类别	模型	内容或行为		社交关系		检测
		图拓扑	特征	图拓扑	特征	实时性
有监督	文献[23]	×	×	√	×	×
	文献[24]	×	×	√	×	×
	文献[26]	×	×	√	×	×
	文献[27]	×	×	√	×	×
	文献[41]	√	×	√	×	×
	文献[28]	√	×	×	×	×
	文献[29]	×	×	√	×	×
	文献[33]	√	×	√	×	×
	文献[35]	×	×	√	×	×
	文献[36]	×	×	√	×	×
	文献[37]	×	×	√	×	×
	文献[38]	×	×	√	×	×
	文献[39]	×	×	√	×	×
	文献[18]	×	√	×	×	×
	文献[1]	×	√	×	×	×
	文献[2]	×	√	×	×	×
	文献[3]	×	√	×	×	×
	文献[4]	×	√	×	×	×
	文献[5]	×	√	×	×	×
	文献[13]	×	√	×	×	×
	文献[14]	×	√	×	×	×
	文献[15]	×	√	×	×	×
	文献[16]	×	√	×	×	×
	文献[17]	×	√	×	×	×
	文献[40]	×	×	×	×	√
无监督	文献[7]	√	×	×	×	×
	文献[19]	×	×	√	×	×
	文献[30]	×	×	√	×	×
	文献[31]	×	×	√	×	×
	文献[32]	√	×	√	×	×
	文献[34]	×	×	√	×	×
	文献[6]	×	√	×	×	×
	文献[11]	×	√	×	×	×
	文献[12]	×	√	×	×	×
	本文方法	×	×	×	×	√

注:“√”表示具备特性;“×”表示不具备特性.

2 恶意注册账号分析

微信作为中国最大的移动互联网社交网络,现已具有 10 亿的月活用户①,背后则是每天百万级别的用户注册量.为了尽早检测出批量注册的恶意账号并防止恶意账号作恶,本文通过分析微信的账号注册数据,挖掘恶意账号的模式与特点,并设计和提出相应的检测算法.本文工作虽基于微信账号注册数据,但本文所分析和总结出的在线社交网络恶意账号在注册阶段所表现出的重要特性,包括恶意账号间的相似特征及与正常用户不同的异常特征,均源自于黑色产业链所运用的自动化批量账号技术,与正常用户的人工注册存在本质不同.故本文所分析的恶意账号注册特性在其他在线社交网络平台上同样适用.据此,本文研究方法通用性好,可进一步运用于其他社交网络场景中的恶意账号检测任务中.

2.1 注册数据

账号注册数据主要包括一系列注册属性,比如注册 IP 地址、昵称、手机号码、WiFi MAC、注册设备 ID、微信客户端版本号、注册设备类型、注册时间、注册国家等.为了保护用户隐私,WiFi MAC、注册设备 ID 等在收集前已经被哈希加密(经哈希处理后,理论上用户和属性值信息仍然能保持一一对应关系,不会对后续分析造成干扰);注册国家、微信客户端版本号等均用代号表示.据抽样观察时间跨度达 3 个月的微信注册账号数据发现,正常账号和恶意账号的比例及各项特征的分布较为稳定.本文进一步对按不同时间跨度划分的注册数据做了验证,结果表明按天划分的数据集的统计结果更具有区分度与代表性.据此,本节将对某一天的微信注册数据进行细致的统计分析.

2.2 注册属性分析

受制于有限的时间资源与设备资源,同时为最大化攻击效率及不法获益,黑色产业链通常会运用自动化账户注册技术等批量获取账号,进而导致在多注册属性上呈现出相似性,且与正常注册账号表现相异.

2.2.1 IP 地址

本文首先对账号注册时使用的 IP 地址前缀进行了统计,统计时采用的前缀长度是 24.结果显示,在正常账号中,IP 地址前缀相同的账号数一般小于 50;而在恶意账号中,IP 地址前缀相同的账号数超过了 50 的情况较多.该现象表明恶意账号比正常账号更倾向使用相同的 IP 前缀进行注册.

① http://www.xinhuanet.com/2018-03/05/c_1122488991.htm

2.2.2　手机号码

注册微信账号需提供手机号码.恶意用户为持续注册账号,需要从通信运营商批量获取手机号.需要注意的是,手机号码的末 4 位是用户的个人编号,而除去末 4 位后的手机号码前缀,则包含有该号码的服务提供商和区域信息.经统计除去末 4 位后的手机号码的前缀发现:所有恶意账号中所使用的手机号码前缀总量较少,单个号码前缀被复用 10 次以上较为常见;所有正常账号中所使用的手机号码前缀总量与正常账号总量接近,单个号码前缀复用 3 次以下的情形较为常见.这表明恶意账号更倾向使用相同的手机号码前缀进行注册.

2.2.3　WiFi MAC

当前手机接入网络主要通过 2 种方式:蜂窝网络或 WiFi.WiFi MAC 是指当移动设备通过 WiFi 接入网络时,WiFi 网关的 MAC 地址.如果注册时使用的是蜂窝网络,则账号的 WiFi MAC 属性为空.本文经统计分析发现,恶意账号更倾向于使用相同的 WiFi MAC 进行注册.

2.2.4　设备 ID

经统计每个设备 ID 所关联的注册账号数量,本文发现共用同一台设备的恶意账号数量远远多于正常账号.此现象表明恶意账号更倾向于使用相同的设备进行批量注册.

2.2.5　昵称模式

本文对正常与恶意账号所使用的昵称进行了以字符粒度的统计分析,发现正常账号昵称通常由中文与个性化字符所组成,而恶意账号昵称往往包含有特殊字符,如冒号、分号等,且恶意账号间往往会共用相同的特殊昵称模式.

2.2.6　客户端版本与操作系统类型

通过对用户注册账号时所使用的微信客户端版本号和手机操作系统类型的分析,本文发现使用老旧客户端或操作系统注册的账号集合中,恶意账号所占比例极大.具体地,数据集中有约 2 000 个账号是基于一个老旧的安卓系统注册的,其中恶意账号占比 96.5%.此外,在 iOS 8(一个老旧的 iOS 操作系统)系统下注册的所有账号中,99%的注册账号是恶意的.此现象背后的原因是黑产出于成本、稳定性等考虑而更倾向于使用老旧的设备与自动化注册脚本.

2.2.7　地理位置

经映射,一个公网 IP 地址可以对应到一个地理位置(国家、省、市),手机号码同理.通过对比注册账号时用户的 IP 地址对应的地理位置与手机号码对应的地理位置,本文发现 65%的恶意账号表现出了地理位置不一致的现象,而正常账号的两地理位置均基本一致.该现象一个可能的原因是,黑产从业者用于注册账号的手机号码可能是从当地的通信运营商处获得的,而用于注册账号的设备可能是远程设备或云服务;另一个可能的原因则是黑产使用的手机号码是从外地购买得到的[42],设备则是本地的.因此,恶意账号注册时更容易出现不一致地理位置的现象.

2.2.8　IP-WiFi 多对多

本文对正常账号和恶意账号在注册时所使用的 IP 和 WiFi MAC 两个属性的对应关系进行了统计分析,发现恶意账号中的单一 WiFi MAC 可能对应着多个 IP,同时这些 IP 又可能对应着多个 WiFi MAC.而在正常账号中,此类型的 IP 与 WiFi MAC 数量稀少.该现象背后的原因是,恶意账号很可能是使用虚拟设备注册的,因而 IP 与 WiFi MAC 之间存在着多对多映射关系.其恰好展现了该类注册账号的虚假性.

2.2.9　注册时间

本文于图 1 与图 2 中展示了不同账号的注册时间分布.其中图 1 是对正常账号统计的结果,图 2 是对恶意账号统计的结果.图 1 和图 2 中不同的折线表示不同的 IP 段.比较发现,正常账号的注册时间分布比较一致,且在半夜仅有很少注册量,与大多数人的生活作息相符.恶意账号的注册时间则分布混乱,不仅均匀分布在 24 h 里,还在某些较短时间内较为密集,与正常账号差异明显.

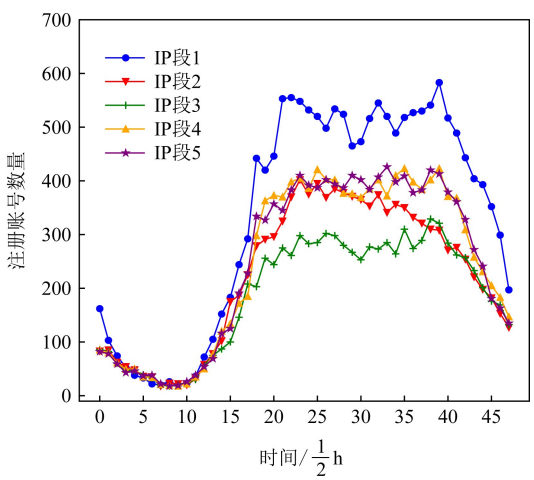


Fig. 1　Distribution of normal account registrations

图 1　正常账号注册分布

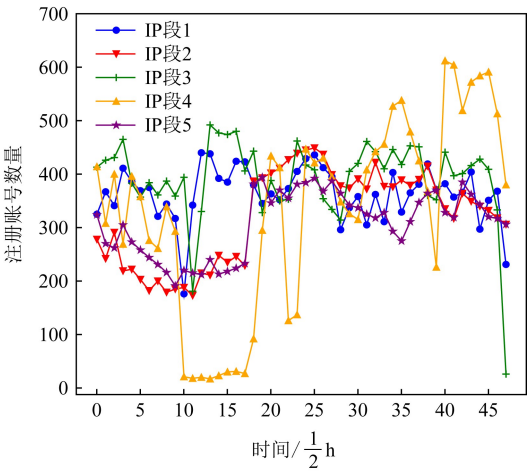


Fig. 2 Distribution of malicious account registrations

图 2 恶意账号注册分布

3 基于注册属性的恶意账号检测

根据第 2 节对账号注册数据的深入统计分析，本文发现恶意账号容易使用某些相同的注册属性，并会在一些属性上与正常注册账户表现相异.这是因为受制于有限的各项资源，恶意团伙一般使用批量注册的方法降低成本，包括使用同一批设备、IP 注册、在相同时间注册等.基于这些特性，本文提出了一种基于注册属性的恶意账号检测方法，其主要运用了无监督图聚类技术.方法具体由 5 个步骤组成：1)账号注册特征提取.2)特征权重配置.3)账号相似度计算.4)账号相似图构建.5)基于图聚类的恶意账号群体挖掘.

在整体上，该方法首先提取各用户账号的注册特征数据.随后，方法基于预配置的权重策略，计算不同账号间的注册属性相似度以构建各注册账号的连通图，并最终通过图聚类方法，挖掘连通图中的特定群体来有效识别恶意账号.方法整体流程如图 3 所示：

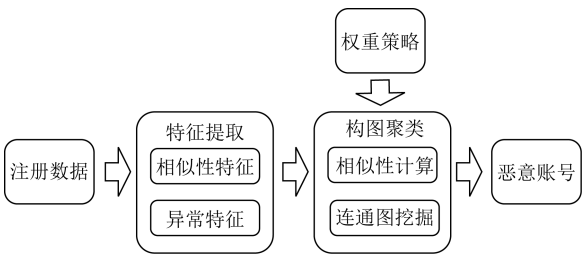


Fig. 3 Overview of our registration pattern based wechat sybils detection method

图 3 基于注册属性的微信恶意账号检测方法流程图

3.1 注册特征提取

根据数据分析阶段的发现，本文首先对各用户账号提取多种注册属性，并基于注册属性进一步获取它们的相似性及异常特征.方法所提取的注册属性及各属性可获取的特征如表 2 所示：

Table 2 Registration Patterns and Features Extracted by Our Method

表 2 本文方法所提取的注册属性及特征

注册属性	相似性特征	异常特征
IP	IP 地址、IP24 位前缀是否相同	是否存在单属性注册数量异常
手机号码	手机号码前缀是否相同	是否存在单属性注册数量异常
WiFi MAC	WiFi MAC 是否相同	是否存在单属性注册数量异常
设备 ID	设备 ID 是否相同	是否存在单属性注册数量异常
昵称模式	昵称模式是否相同	是否为异常的昵称模式
客户端版本号	客户端版本号是否相同	是否为老旧的客户端版本号
操作系统类型	操作系统类型是否相同	是否为老旧操作系统类型
地理位置		是否存在地理位置异常
IP-WiFi 多对多		是否存在 IP-WiFi 多对多异常
注册时间		注册时间分布是否异常；是否半夜注册

3.1.1 昵称模式

不同于其他可直接提取的注册属性，本文基于特定的昵称模式及提取方法来更针对性处理各账号的注册昵称.根据字符特征，昵称可分为中文、英文、混合字符 3 种类型.从中文昵称中，可以提取出传统中文姓名和随机中文字符串 2 种昵称模式；从英文昵称中，可以提取出拼音和随机英文字符串 2 种昵称模式；对于混合字符昵称，将其抽象化后得到的符号串作为其昵称模式.具体针对不同类型的昵称，本文采用这 3 种昵称模式提取方法：

1) 中文昵称.本文将中文昵称分为传统中文姓名和随机中文字符串.①传统中文姓名.使用结巴分词工具①对中文昵称进行分词,然后在百家姓的基础上进行一次过滤,再根据词性分析以及字符串长度过滤,2 次过滤后得到的所有中文昵称的昵称模式就是传统中文姓名.②随机中文字符串.使用正常的中文文章集作为语料库训练 n-gram 模型②.如果一个字符串是中文常用语,将其输入到训练好的 n-gram 模型中,模型会输出一个较高的概率值,表明模型认为该字符串出现在语料库中的可能性较高;反之则是低概率值,表明该字符串出现在语料库中的可能性低,即是随机的中文字符串.为了确定模型阈值,本文从昵称数据集中抽取少量已知是随机中文字符串的昵称(本文抽取了 1 000 个昵称),将其输入 n-gram 模型计算概率值,并根据这些随机中文昵称的概率值的分布来启发式地设置了阈值.如果输入字符串的概率值低于该阈值,则认为输入昵称为随机中文字符串昵称.

2) 英文昵称.本文将英文昵称分为拼音和随机英文字符串 2 种模式.如同提取中文昵称模式时所述的方法,本文分别使用对应的语料库来训练正常拼音或英文字符串的 n-gram 模型.根据随机抽取的拼音昵称或随机英文字符串昵称的概率值,启发式地设置了阈值使得模型能检测出拼音昵称或随机英文字符串昵称.

3)混合字符.混合字符一般具有较为明显的模式,比如中英混合、夹杂特殊符号(比如分号、@)等.本文采用了 Thomas 等人[18]所提出的方法,以将昵称抽象化成统一类型的符号串.具体地,本文首先设计一类字符映射规则,将昵称中不同类型的字符映射到不同的符号上.比如中文字符用 C 表示,大写字母字符用 U 表示,小写字母字符用 L 表示,数字字符用 D 表示,其余字符保留原始字符.在该映射规则下,“张三 123”将被映射为“CCLLL”.映射后得到的符号串,就是昵称的模式,即符号串“CCLLL”就是昵称“张三 123”的昵称模式.

3.1.2　相似性特征

基于各账号的注册属性,本文提取 8 个相似性特征来判断 2 个账号是否在某些角度表现出相似性:

1) IP 前缀相同.如果 2 个账号的 24 位 IP 前缀相同,则 2 个账号具有该特征.

2) IP 相同.如果 2 个账号的 IP 地址完全相同,则 2 个账号具有该特征.

3) 手机号码前缀相同.如果 2 个账号的手机号码前缀相同,则 2 个账号具有该特征.

4) WiFi MAC 相同.如果 2 个账号的 WiFi MAC 相同,则 2 个账号具有该特征.

5) 设备 ID 相同.如果 2 个账号的设备 ID 相同,则 2 个账号具有该特征.

6) 客户端版本号相同.如果 2 个账号的微信客户端版本号相同,则 2 个账号具有该特征.

7) 设备操作系统类型相同.如果 2 个账号的注册设备系统类型相同,则 2 个账号具有该特征.

8) 昵称模式相同.若 2 个账号具有相同的昵称模式,如均为中文名字昵称、拼音昵称或两者昵称抽象模式之间的编辑距离与其长度平均值的比例小于一定阈值(基于本文对注册账号昵称数据的统计分析,将该阈值设为 0.3),则认为 2 个账号具有该特征.

3.1.3　异常特征

本文提取的异常特征则包含 9 个部分:

1) 老旧客户端版本号.本文根据当前已发布的最新微信版本,来确定老旧的或者罕见的微信客户端版本号.若 2 个账号客户端均老旧,则认为它们具有该特征.

2) 老旧设备系统类型.本文根据各个手机操作系统的更新历史,来确定老旧的或者罕见的设备系统类型.若 2 个账号操作系统均老旧,则认为它们具有该特征.

3) 单属性注册数量.本文同样通过统计分析数据来制定判断单个相同属性下注册账号数量是否超过阈值及是否可认定为异常.具体地,若同一个 IP 地址下注册的账号超过 40 个,或同一个 WiFi MAC/设备 ID 下注册的账号超过 25 个,或同一个手机号前缀下超过 30 个,则认定这个 IP 地址、WiFi MAC、设备 ID 或者手机号前缀是异常的;若 2 个账号均具有单属性注册数量异常,则认为它们存在对应特征.

4) 地理位置异常.若 2 个账号手机号码前缀相同,但 IP 不同,则本文认定它们为异常.若 IP 相同,但手机号码前缀不同,则同样认定 2 个账号是异常的;若 2 个账号均具有该异常属性,则认为它们具有地理位置异常特征.

5) IP-WiFi 多对多异常.倘若 IP 与 WiFi 存在

① Jieba. https://github.com/fxsjy/jieba

② Wikipedia contributors. N-gram. Wikipedia, The Free Encyclopedia. April 11, 2018. https://en.wikipedia.org/w/index.php?title=N-gram&oldid=835900923. Accessed June 29, 2018

多对多关系,即一个 WiFi MAC 对应多个 IP.同时,被对应的 IP 也对应着多个 WiFi MAC,则本文认为这样的 IP 和 WiFi MAC 是异常的;若 2 个账号均具有 IP-WiFi 多对多异常现象,则认为它们存在 IP-WiFi 多对多异常特征.

6) 时间分布异常.本文将一个 IP 段下的账号注册时间分布与正常账号的注册时间分布进行对比,如果 KL 距离大于设定阈值,那么认定该 IP 段是异常的.本文同样基于数据分析的结果启发式地调整该阈值,最终使用默认值 1.0 作为 KL 距离的阈值;若 2 个账号均具有时间分布异常属性,则认为它们存在时间分布异常特征.

7) 昵称模式异常.若 2 个账号的昵称模式均为随机中文字符串或者随机英文字符串,则本文认为 2 个账号在该特征上异常.

8) 注册国家异常.若用户填写的注册国家和其真实注册的国家不一致,则该账号在注册国家上是异常的;若 2 个账号均存在该现象,则认为它们存在注册国家异常特征.

9) 注册时间异常.如果 2 个账号均在半夜进行注册(本文使用的是在凌晨 2:00—5:00),则本文认为该 2 个账号存在注册时间异常特征.

3.2 构图聚类

本文提出了一种基于构图聚类的恶意注册账号检测方法.该方法基于各账号注册属性及提取的相似性和异常特征,视各账号为图中顶点,计算节点间相似性并建立边,从而构建一张账号间的相似性图.随后,通过进行图聚类和计算不同账号的恶意分数,从而准确识别图中的恶意账号.

3.2.1 特征权重分配

考虑到不同的特征在决策账号间相似性时作用可能存在不同,本文首先为每个特征赋予一定的初始权值,其由对数据中相似恶意账号的统计分析结果来进行启发式地确定.本文将特征的重要性分为 4 个等级,从低到高依次设为 0.5,1.0,1.5,2.0.例如版本号相似、操作系统相似等恶意账号间常见的特征,本文为它们配置最低初始权重.例如 WiFi MAC 相似、IP 相似等包含地理位置等信息的特征,本文为它们配置最高初始权重.

在初始权重的基础上,本文进一步通过多轮实验迭代的方式进行权重调整,以期更好地表现出微信数据中恶意注册账号的特点.本文通过在 3.1.2 节和 3.1.3 节所述分析数据上的构图聚类检测结果,且在每轮迭代时只改变一个特征的权重等级,进而对比上一轮实验和本次实验结果的优劣,最终只记录能提升性能的权重特征改变.如此对全部特征权重进行逐个调整,并最终形成如表 3 所示的特征权重表,以用于对测试数据的实际检测中.

Table 3 Weights of Our Features

表 3 特征权重分配

类别	特征	权重值
相似性特征	IP 前缀相同	1.0
	IP 相同	0.5
	手机号码前缀相同	1.0
	WiFi MAC 相同	1.5
	设备 ID 相同	2.0
	客户端版本号相同	0.5
	设备操作系统类型相同	0.5
	昵称模式相同	1.0
异常特征	老旧客户端版本号	1.0
	老旧设备操作系统类型	1.0
	单属性注册数量异常	1.0
	地理位置异常	0.5
	IP-WiFi 多对多异常	1.0
	时间分布异常	0.5
	昵称模式异常	1.0
	注册时间异常	0.5
	注册国家异常	0.5

3.2.2 边的权重计算与连通图建立

基于 3.1.2 节、3.1.3 节和 3.2.1 节所述的特征与特征权重,本文首先给出账号间注册相似性($Similarity$)的定义:

定义 1. 账号相似性 $Similarity$.设 C_1 与 C_2 为全体账号集合 C 中任意 2 个账号,CF_1 与 CF_2 分别为 C_1 与 C_2 对应的注册属性,$Feat(x,y)$ 为基于 2 个账号各自注册属性进行 3.1.2 和 3.1.3 节所述账号间特征提取函数,$\boldsymbol{f}$ 则为 3.2.1 节所述特征权重向量,则 C_1 与 C_2 间的账号相似性 $Similarity$ 的计算方式为

$$Similarity(C_1,C_2)=\boldsymbol{f}\cdot Feat(CF_1,CF_2),\quad(1)$$

其中"·"为内积运算符.

基于账号间注册相似性的定义,本文进一步给出账号相似连通图的定义:

定义 2. 账号相似连通图 G.设账号相似连通图 $G=(V,E)$,其中 V 为各注册账号所组成的顶点集合.对于 $\forall V_1,V_2\in V$,若 $Similarity(V_1,V_2)>$

$Threshold$,则 E 中存在且唯一存在一条边 $e=(V_1,V_2)$.另设权重 $W(e)=Similarity(V_1,V_2)$.

对于决定是否在 2 个顶点间建边的相似度阈值 $Threshold$,本文同样通过实验迭代的方式进行设置.本文最终所使用的相似度阈值为 3.5.

3.2.3　连通图聚类

基于批量注册的恶意账号具备较强相似性这一特性,本文通过采用图聚类的方式从相似连通图中挖掘出由相似顶点组成的簇.具体地,本文采用典型图聚类方法:对建立好的账号间相似关系图(非连通无向图)进行遍历,获取图中所有连通分量.每个连通分量都是相似连通图的一个极大连通子图,从连通分量中的任一顶点出发,能且只能访问到该连通分量中所有的顶点.将同一个连通分量内的所有注册账号归为一个账号簇,其中所有账号间表现出较强的相似性.

然而,特定部分正常账号可能与恶意账号存在较为相似的属性,进而被关联至特定账号簇中.故本文将对各账号簇进行更细致的分析,从而准确地检测出簇中的恶意账号.

3.2.4　恶意分数计算

为了避免将特定正常账号误判为恶意账号,本文基于连通分量中各顶点的边权重来进行更细致的判断.其依据是相较恶意注册账号本身,特定正常账号与恶意账号间存在边(相似)的可能性更少,故它们的边数与权重更少.据此,本文计算每个账号的恶意分数 Malicious Score 来判断.

定义 3. 账号恶意分数 Malicious Score.给定账号连通图 $G=(V,E)$,对于 $\forall v\in V$,设以其作为顶点的边集合为 $Edge(v)$,则定义 v 的 Malicious Score 的计算方式为

$$MaliciousScore(v)=\tanh\Big(\sum_{i=1}^{|Edge(v)|}W(e_i)\Big),\quad e_i\in Edge(v). \tag{2}$$

对于每个挖掘出的连通分量中的每个账号顶点,本文计算其所连接边权重和的 tanh 值,以表示该账号的恶意分数.恶意分数越高,则代表该账号与其他账号的相似性越大,亦代表该账号为恶意的可能性越大.计算出各账号的恶意分数后即可通过阈值对比来最终判断各账号是否为恶意.该阈值采用了与特征权重相同的方式,在分析数据上进行配置和调整,并最终用于测试数据的检测中,该阈值最终设为 0.75.

4　性能评价

我们使用 Scala 编程语言实现了基于本文方法的原型系统,并基于 Spark 框架实现百万级别注册账号的快速准确检测能力.该系统已在微信应用平台进行了部署和较长时期应用.

4.1　实验数据集

本文实验使用的数据为微信应用 2017 年 10 月某一周的用户账号注册日志.数据总计有 1 040 万条注册记录,其中恶意账号数量为 500 万.平均每一天约有 150 万的注册账号,其中恶意账号占比最高为 50%左右.

1) 数据集划分.一周的总数据集按天被分成 7 份.本文使用第 1 天的分析数据确定了表 3 所示的各特征权重、相似度阈值(值为 3.5)和恶意分数阈值(值为 0.75).这些权重和阈值是通过启发式的方法,根据模型在分析数据上的实验结果不断地调整而确定的.测试时,分别取第 1 天、前 3 天、前 5 天、前 7 天的数据用于验证和对比模型的检测效果.

2) 数据集标签.为了验证模型的检测效果,腾讯公司提供了数据集所有数据的标签.标签来源于其他根据注册后用户的行为来检测的相关模型、其他用户的举报以及微信安全团队的抽样审计.

3) 实验环境.本文系统构建于腾讯公司 Spark 计算平台之上.Spark 是业内常用的大规模数据计算引擎.在实验中,本文系统共使用了 30 个 Executor,每个 Executor 配置了 16 核 CPU 和 10 GB 内存.

4) 用户隐私保护.本文工作是与微信安全团队的合作项目,已签署了微信的相关保密协议.实验使用的数据源于用户注册账号时收集的相关信息.所有收集的数据都已在微信的隐私保护协议中声明,用户在注册账号前必须阅读且同意该协议才可以注册账号.在收集数据时,微信会对敏感数据先脱敏再收集.如用户手机的号码仅保留号码前缀,对 WiFi MAC 和设备 ID 进行哈希计算等.

4.2　系统实现

在通过 Spark 平台实现基于本文方法的原型系统时,考虑到日均 150 万的新增注册账号数量,直接基于任意一对账号的注册属性获取二者之间的相似性与异常特征,进而计算二者相似度,最终构建账号相似度连通图这样的实现方法计算开销将是巨大而不可接受的.据此,本文采用了如下工程方式来有效加速账号间相似度计算,从而使本文原型系统能满足微信应用日均百万级的新注册账号检测需求:

在读入当日新账号的每个注册属性时,就对所有账号按相同的单个属性值进行划分,即在各注册属性上将全体账号划分为多个集合,每个集合包含一定数量的账号.由于本文设计的相似性和异常特征均提取自含有相同属性值的账号对,故只需对各集合内部的账号对进行特征提取与相似度计算,进而遍历当前注册属性的所有集合及其他所有注册属性的账号划分集合,即可完成所有可能存在相似性的账号对的计算,而无需穷举全体账号内的任意账号对.

设全体账号数目为 N,计算任一账号对某一特征的时间开销为 $O(1)$,特征数目为 m,则直接计算全体账号间整体相似度的时间开销为 $O(mN^2)$.设对于全体账号每个注册属性均存在 a 个不同值,每个值下对应了 $\frac{N}{a}$ 个账号,则基于同属性值划分计算所有可能存在相似关系的账号节点时间开销为 $O\left(\left(\left(\frac{N}{a}\right)^2\times m\right)\times a\times m\right)$,即为 $O\left(\frac{(mN)^2}{a}\right)$,其中 $O\left(\left(\frac{N}{a}\right)^2\times m\right)$ 为基于某个注册属性某个属性值对应的账号集合内部进行账号间相似度计算(对全体特征)的时间开销,$a\times m$ 则为遍历该注册属性所有属性值对应集合及所有注册属性(特征)所需系数.由于当前特征数仅为 17(8 个相似性特征,9 个异常特征),在绝大多数情况下各属性可能拥有值的数目将远远大于特征数,即 $a\gg m$,故 $O\left(\frac{(mN)^2}{a}\right)\ll O(mN^2)$.由此可见,本文的工程化系统实现方法能显著加速账号连通图的构建过程,从而保障系统满足微信应用的日常新注册账户检测需求.

此外,该实现方法的本质是基于相同注册属性先选出在特定特征的部分账号对(可能存在边),忽略其余明确不存在特定特征的账号对(肯定不存在边),进而再基于挑选出的账号对进行相似度计算(确定是否建边).据此,此种方法下账号连通图中的边与原始方法下连通图中的边将不会存在差异,对连通分量的挖掘与恶意账号检测结果同样不会受到任何影响.

4.3 准确率和召回率

若对恶意账号的检测结果中存在较多的误报,则容易造成大量正常用户被误封,将严重影响微信应用的用户体验和正常运行.据此,为确保检测出恶意账号的准确性,本文选择了较高的相似度阈值,以在可接受的召回率前提下使检测结果尽可能准确.如图 4 所示,本文方法检测结果的准确率为 96%左右,召回率为 50%~60%,随着数据量从百万级增长到千万级,本文的算法仍然保持相对稳定的性能.图 5 则展示了各测试数据集下,本文方法的日平均检测结果,每天平均可准确检测出 40 万至 50 万的恶意注册账号.

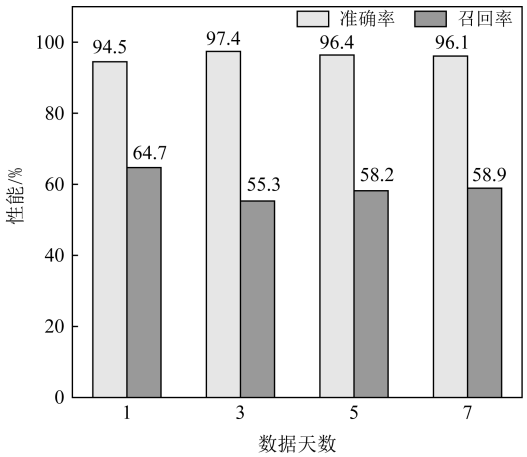


Fig. 4 Precision and recall of our method on the dataset collected in Oct. 2017

图 4 基于 2017 年 10 月数据集的准确率和召回率

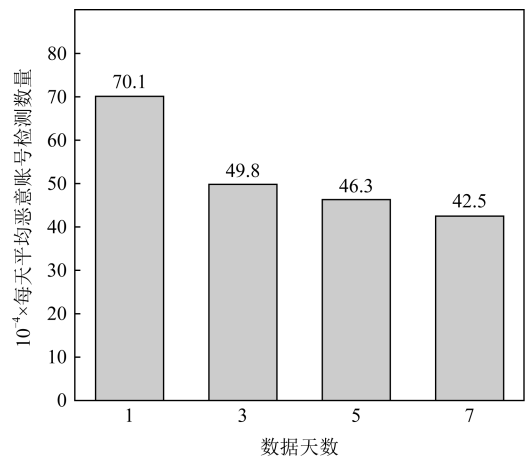


Fig. 5 Average number of detected sybils per day

图 5 每天平均的恶意账号检测数量

4.4 时效性

为了验证模型是否能在一段时间之后的数据集上仍保持较好的效果,本文又获取了微信应用 2017 年 11 月某一周连续 5 天的用户账号注册日志.数据总计有 690 万条注册记录,其中恶意账号数量为 294 万.使用与 4.2 节中同样的参数和阈值,模型对每天注册账号检测的准确率和召回率如图 6 所示.对比 4.3 节的模型检测结果,可以看到在新数据集上模型有 92%左右的准确率和 70%左右的召回率,

依然保持着稳定的检测效果.由此可见,本文方法及选定阈值并不局限于特定时期的数据,具有良好的时效性.

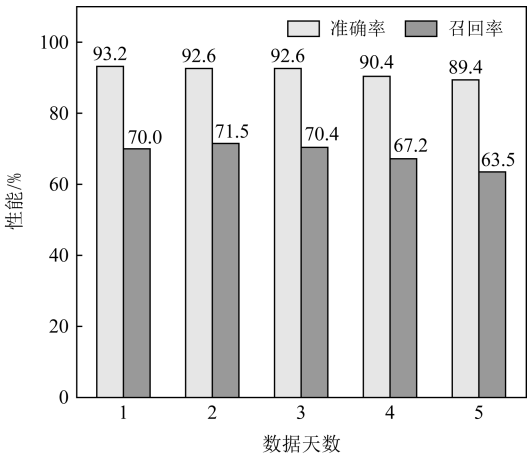


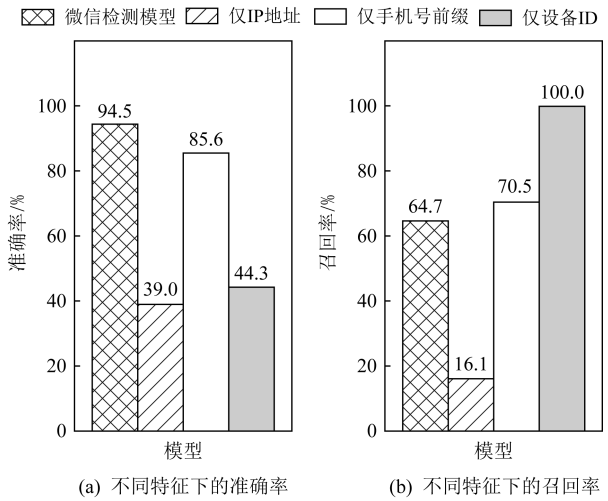
Fig. 6　Precision and recall of our method on the dataset collected in Nov. 2017

图 6　基于 2017 年 11 月数据集的准确率和召回率

4.5　特征重要性

根据第 2 节的注册数据分析结果,在直观上我们可以提出更加朴素的检测方法:即仅基于 IP 地址或手机号等注册属性来进行检测,而不再需要 3.1 节所设计的相似性特征和异常特征.为了验证本文设计的相似性特征和异常特征的重要性,本文进行了一个对比实验.以本文提出的微信恶意账号检测模型为基准线,分别与仅用 IP 地址、仅用手机号前缀和仅用设备 ID 三种不同检测模型进行对比.本文仅选用 IP 地址、手机号和设备 ID 这 3 个属性的原因是它们对恶意和正常账号的区分度远优于其他属性(如昵称等).以仅用 IP 地址的检测模型为例,根据分析数据中同一恶意 IP 地址下注册的账号数与同一正常 IP 地址下注册的账号数,选择一个合适的阈值(本文使用了 500),如果测试数据中一个 IP 地址下注册的账号数超过该阈值,则此 IP 地址下的所有账号都是恶意账号,反之此 IP 地址下的所有账号都是正常账号.在仅用手机号前缀的检测模型中,本文使用的阈值是 9;在仅用设备 ID 的检测模型中,本文使用的阈值是 2.这些阈值是根据各模型的检测结果启发地调整得到的,使得各模型能够取得最优的检测结果.需要注意的是,因为是仅用单个属性来检测,所以阈值与 3.1.3 节单属性注册异常中的阈值不一样.对比实验的结果如图 7 所示.

从图 7 中可以看到,相比于本文提出的微信恶意账号检测模型,仅用单个属性的朴素模型在准确率上有明显的差距.这就意味着,在实际使用场景中,这些模型会造成大量用户的账号被误封,极大影响用户体验.这个对比实验表明,本文提取的相似性或异常特征是重要的,可以保证模型对恶意账号覆盖率和提高检测结果的准确率.



(a) 不同特征下的准确率　(b) 不同特征下的召回率

Fig. 7　Precision and recall of our method under different feature sets

图 7　不同特征下的准确率与召回率实验结果

4.6　可拓展性

基于实际场景测试,本文所实现原型系统检测百万级别的数据需耗时约 20 min,检测千万级别的数据则需耗时约 100 min,具体如图 8 所示.可以看出,本文所实现原型系统对于千万级的数据仍然有较快的检测速度,并且检测结果的准确率稳定在 96%左右.由此可见,本文方法及原型系统可拓展性良好.

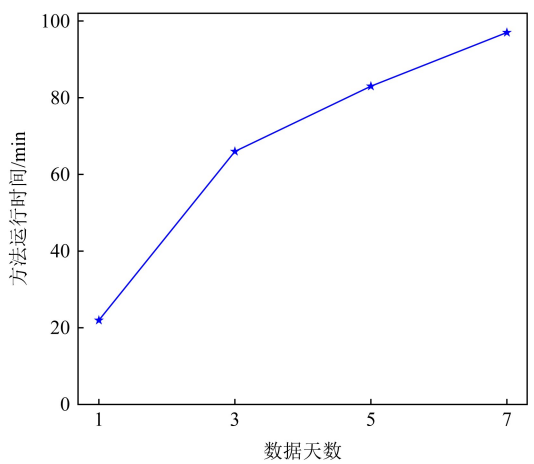


Fig. 8　Running time of our method on different test data

图 8　不同测试数据下方法的运行时间

4.7 系统部署

本文提出的无监督恶意注册检测模型及原型系统已被微信安全团队部署,用于对恶意注册账号的日常检测.具体地,检测模型会于每天 24 时,基于账号注册日志对当天的所有注册账号进行线下检测.由于模型存在着少量的误报进而导致误封号,而这些被误封号的用户可以通过微信内置的解封申诉渠道进行解封.由此产生的误封记录将被维护人员分析,用于进一步地提升模型的检测效果.据近期统计,本文系统每日依然可检测出 40 万左右的恶意账号,根据用户的申诉反馈和维护人员的抽样分析,检测准确率在 90%~95%.需要注意的是,微信公司每日新增注册用户数量在百万级,故无法知悉其中全部恶意账号数目并计算本文系统实际部署阶段的漏报率或召回率.

异常检测系统在实际部署中同样需要考虑先进攻击者进行针对性逃避的可能.对于目前比较先进的黑产攻击者,他们的主要表现是注册成功后各类攻击行为更加复杂隐蔽.而本文方法瞄准的是其注册阶段,即在其发动攻击之前进行有效防范.本文方法所使用的注册属性在新用户注册账号时必须提交,且会由于自动注册工具的使用而表现出聚集性.典型的黑产攻击者均无法回避这 2 个特性.因此,本文方法能有效防范真实网络中大部分典型的黑产攻击者威胁.

5 总 结

本文首先系统介绍并分析了现有社交网络恶意账号检测方法的原理及优缺点.随后,通过全面地对社交网络账号的注册属性进行统计分析,对比正常和恶意账号在不同注册属性上的分布差异,本文设计了相似性特征和异常特征用于比较不同注册账号间的相似性,进而构建账号相似连通图,并通过连通图算法挖掘并检测恶意注册账号.本文方法具备不依赖带标签训练集、检测性能好且稳定、处理速度快等优点,并已在微信运营平台得到了实际部署和长期应用,有力打击了在线社交网络黑产链,保障广大用户安全使用微信应用.

参 考 文 献

[1] Almaatouq A, Shmueli E, Nouh M, et al. If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts [J]. International Journal of Information Security, 2016, 15(5): 475-491

[2] Egele M, Stringhini G, Kruegel C, et al. Towards detecting compromised accounts on social networks [J]. IEEE Transactions on Dependable and Secure Computing, 2015, 14(4): 447-460

[3] Freeman D, Jain S, Dürmuth M, et al. Who Are You? A statistical approach to measuring user authenticity [C] //Proc of the 23rd Annual Network and Distributed System Security Symp. Reston, VA: ISOC, 2016: 1-15

[4] Badri Satya P R, Lee K, Lee D, et al. Uncovering fake likers in online social networks [C] //Proc of the 25th ACM Int on Conf on Information and Knowledge Management. New York: ACM, 2016: 2365-2370

[5] Wang Gang, Konolige T, Wilson C, et al. You are how you click: Clickstream analysis for sybil detection [C] //Proc of the 22nd USENIX Security Symp. Berkeley, CA: USENIX Association, 2013: 241-256

[6] Cao Qiang, Yang Xiaowei, Yu Jieqi, et al. Uncovering large groups of active malicious accounts in online social networks [C] //Proc of the 2014 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2014: 477-488

[7] Jiang Meng, Cui Peng, Beutel A, et al. Catchsync: Catching synchronized behavior in large directed graphs [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 941-950

[8] Boshmaf Y, Logothetis D, Siganos G, et al. Integro: Leveraging victim prediction for robust fake account detection in OSNs [C] //Proc of the 22nd Annual Network and Distributed System Security Symp. Reston, VA: ISOC, 2015: 8-11

[9] Yu Haifeng, Gibbons P B, Kaminsky M, et al. Sybillimit: A near-optimal social network defense against sybil attacks [J]. IEEE/ACM Transactions on Networking, 2010, 18(3): 885-898

[10] Cao Qiang, Sirivianos M, Yang Xiaowei, et al. Aiding the detection of fake accounts in large scale social online services [C] //Proc of the 9th USENIX Conf on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2012: 197-210

[11] Stringhini G, Mourlanne P, Jacob G, et al. Evilcohort: Detecting communities of malicious accounts on online services [C] //Proc of the 24th USENIX Security Symp. Berkeley, CA: USENIX Association, 2015: 563-578

[12] Gao Hongyu, Hu Jun, Wilson C, et al. Detecting and characterizing social spam campaigns [C] //Proc of the 10th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2010: 35-47

[13] Song J, Lee S, Kim J. Spam filtering in Twitter using sender-receiver relationship [C] //Proc of the 14th Int Workshop on Recent Advances in Intrusion Detection. Berlin: Springer, 2011: 301-317

[14] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks [C] //Proc of the 26th Annual Computer Security Applications Conf. New York: ACM, 2010: 1-9

[15] Wang A H. Don't follow me: Spam detection in Twitter [C] //Proc of the 7th Int Conf on Security and Cryptography. Piscataway, NJ: IEEE, 2010: 1-10

[16] Yang Chao, Harkreader R C, Gu Guofei. Die free or live hard? empirical evaluation and new design for fighting evolving Twitter spammers [C] //Proc of the 14th Int Workshop on Recent Advances in Intrusion Detection. Berlin: Springer, 2011: 318-337

[17] Leontjeva A, Goldszmidt M, Xie Yinglian, et al. Early security classification of skype users via machine learning [C] //Proc of the 2013 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2013: 35-44

[18] Thomas K, McCoy D, Grier C, et al. Trafficking fraudulent accounts: Therole of the underground market in Twitter spam and abuse [C] //Proc of the 22nd USENIX Security Symp. Berkeley, CA: USENIX Association, 2013: 195-210

[19] Alvisi L, Clement A, Epasto A, et al. Sok: The evolution of sybil defense via social networks [C] //Proc of the 34th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2013: 382-396

[20] Cai Zhuhua, Jermaine C. The latent community model for detecting sybils in social networks [C] //Proc of the 19th Annual Network and Distributed System Security Symp. Reston, VA: ISOC, 2012

[21] Danezis G, Mittal P. Sybilinfer: Detecting sybil nodes using social networks [C] //Proc of the 16th Annual Network and Distributed System Security Symp. Reston, VA: ISOC, 2009: 1-15

[22] Gong N Z, Frank M, Mittal P. Sybilbelief: A semi-supervised learning approach for structure-based sybil detection [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 976-987

[23] Liu Changchang, Gao Peng, Wright M, et al. Exploiting temporal dynamics in sybil defenses [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 805-816

[24] Mohaisen A, Hopper N, Kim Y. Keep your friends close: Incorporating trust into social network-based sybil defenses [C] //Proc of the 30th IEEE Int Conf on Computer Communications. Piscataway, NJ: IEEE, 2011: 1943-1951

[25] Mohaisen A, Yun A, Kim Y. Measuring the mixing time of social graphs [C] //Proc of the 10th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2010: 383-389

[26] Thomas K, Li F, Grier C, et al. Consequences of connectivity: Characterizing account hijacking on Twitter [C] //Proc of the 2014 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2014: 489-500

[27] Viswanath B, Post A, Gummadi K P, et al. An analysis of social network-based sybil defenses [J]. ACM SIGCOMM Computer Communication Review, 2010, 40(4): 363-374

[28] Yang Chao, Harkreader R, Zhang Jialong, et al. Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 71-80

[29] Xue Jilong, Yang Zhi, Yang Xiaoyong, et al. Votetrust: Leveraging friend invitation graph to defend against social network sybils [C] //Proc of the 32nd IEEE Int Conf on Computer Communications. Piscataway, NJ: IEEE, 2013: 2400-2408

[30] Yu Haifeng, Kaminsky M, Gibbons P B, et al. Sybilguard: Defending against sybil attacks via social networks [J]. ACM SIGCOMM Computer Communication Review, 2006, 36(4): 267-278

[31] Zhao Yao, Xie Yinglian, Yu Fang, et al. BotGraph: Large scale spamming Botnet detection [C] //Proc of the 6th USENIX Symp on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2009, 9: 321-334

[32] Zheng Haizhong, Xue Minhui, Lu Hao, et al. Smoke screener or straight shooter: Detecting elite sybil attacks in user-review social networks [J]. arXiv preprint arXiv:1709.06916, 2017

[33] Yang Zhi, Wilson C, Wang Xiao, et al. Uncovering social network sybils in the wild [J]. ACM Transactions on Knowledge Discovery from Data, 2014, 8(1): 1-29

[34] Xie Yinglian, Yu Fang, Ke Qifa, et al. Innocent by association: Early recognition of legitimate users [C] //Proc of the 2012 ACM Conf on Computer and Communications Security. New York: ACM, 2012: 353-364

[35] Gao Peng, Wang Binghui, Gong N Z, et al. Sybilfuse: Combining local attributes with global structure to perform robust sybil detection [C] //Proc of the 6th IEEE Conf on Communications and Network Security. Piscataway, NJ: IEEE, 2018: 1-9

[36] Jia Jingyuan, Wang Binghui, Gong N Z. Random walk based fake account detection in online social networks [C] //Proc of the 47th Annual IEEE/IFIP Int Conf on Dependable Systems and Networks. Piscataway, NJ: IEEE, 2017: 273-284

[37] Wang Binghui, Gong N Z, Fu Hao. GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs [C] //Proc of the 17th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2017: 465-474

[38] Wang Binghui, Jia Jinyuan, Gong N Z. Graph-based security and privacy analytics via collective classification with joint weight learning and propagation [J]. arXiv preprint arXiv: 1812.01661, 2018

[39] Wang Binghui, Zhang Le, Gong N Z. SybukSCAR: Sybil detection in online social networks via local rule based propagation [C] //Proc of the 36th IEEE Int Conf on Computer Communications. Piscataway, NJ: IEEE, 2017: 1-9

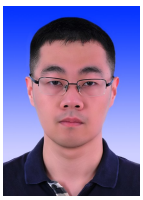
[40] Yuan Dong, Miao Yuanli, Gong N Z, et al. Detecting fake accounts in online social networks at the time of registrations [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 1423-1438

[41] Xia Zenghua, Liu Chang, Gong N Z, et al. Characterizing and detecting malicious accounts in privacy-centric mobile social networks: A case study [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2019: 2012-2022

[42] Thomas K, Huang D, Wang D, et al. Framing dependencies introduced by underground commoditization [C] //Proc of the 14th Annual Workshop on the Economics of Information Security. The Netherlands: Delft, 2015: 1-24

Yang Zheng, born in 1993. Master. His main research interests include machine learning and information security.

杨 征,1993 年生.硕士.主要研究方向为机器学习和信息安全.



Yin Qilei, born in 1991. PhD. His main research interest is cyber security.

殷其雷,1991 年生.博士.主要研究方向为网络安全.

Li Haoran, born in 1994. Master. His main research interests include machine learning and information security.

李浩然,1994 年生.硕士.主要研究方向为机器学习和信息安全.

Miao Yuanli, born in 1993. Master. Her main research interest is cyber security.

苗园莉,1993 年生.硕士.主要研究方向为网络安全.

Yuan Dong, born in 1993. Master. His main research interests include machine learning and information security.

元 东,1993 年生.硕士.主要研究方向为机器学习和信息安全.

Wang Qian, born in 1980. PhD, professor, PhD supervisor. His main research interests include AI security, data storage and search, computation outsourcing security and privacy protection, big data security and privacy, and applied cryptography.

王 骞,1980 年生.博士,教授,博士生导师.主要研究方向为人工智能安全、数据存储与查询、计算外包安全与隐私保护、大数据安全与隐私以及应用密码学.

Shen Chao, born in 1985. PhD, professor, PhD supervisor. His main research interests include cyber-physical system optimization and security, network and system security, and artificial intelligence security.

沈 超,1985 年生.博士,教授,博士生导师.主要研究方向为信息物理融合系统优化与安全、网路和系统安全,以及人工智能安全.

Li Qi, born in 1979. PhD, associate professor, PhD supervisor. Senior member of CCF. His main research interests include network and system security, Internet security, mobile security, and big data security.

李 琦,1979 年生.博士,副教授,博士生导师,CCF 高级会员.主要研究方向为网络和系统安全、互联网安全、移动安全和大数据安全.